

MAIN PAPER

A marginalized two-part joint model for a longitudinal biomarker and a terminal event with application to advanced head and neck cancers

Denis Rustand^{1,2}  | Laurent Briollais³ | Virginie Rondeau¹

¹UMR1219, Inserm, Bordeaux Population Health Center, University of Bordeaux, Bordeaux, France

²Statistics Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

³Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital and Dalla Lana School of Public Health (Biostatistics), University of Toronto, Toronto, Ontario, Canada

Correspondence

Denis Rustand, Statistics Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.
Email: denis@rustand.fr

Funding information

Canadian Institutes of Health Research; Canadian Statistical Science Institute; École des Hautes Études en Santé Publique - Réseau Doctoral; Institut National Du Cancer, Grant/Award Number: 2017-0680; Natural Sciences and Engineering Research Council of Canada

Abstract

The sum of the longest diameter (SLD) of the target lesions is a longitudinal biomarker used to assess tumor response in cancer clinical trials, which can inform about early treatment effect. This biomarker is semicontinuous, often characterized by an excess of zeros and right skewness. Conditional two-part joint models were introduced to account for the excess of zeros in the longitudinal biomarker distribution and link it to a time-to-event outcome. A limitation of the conditional two-part model is that it only provides an effect of covariates, such as treatment, on the conditional mean of positive biomarker values, and not an overall effect on the biomarker, which is often of clinical relevance. As an alternative, we propose in this article, a marginalized two-part joint model (M-TPJM) for the repeated measurements of the SLD and a terminal event, where the covariates affect the overall mean of the biomarker. Our simulation studies assessed the good performance of the marginalized model in terms of estimation and coverage rates. Our application of the M-TPJM to a randomized clinical trial of advanced head and neck cancer shows that the combination of panitumumab in addition with chemotherapy increases the odds of observing a disappearance of all target lesions compared to chemotherapy alone, leading to a possible indirect effect of the combined treatment on time to death.

KEYWORDS

conditional two-part, joint model, left-censoring, marginalized two-part, randomized clinical trial, semicontinuous, solid tumors

1 | INTRODUCTION

In solid tumor cancer clinical trials, there is an increased interest in the joint analysis of the time to death and the sum of the longest diameter (SLD) of the target lesions, defined according to the Response Evaluation Criteria in Solid Tumours (RECIST). This biomarker reflects the tumor burden and its evolution over time. It is important to account for the association between the longitudinal outcome and the risk of terminal event because the former is censored by the terminal event, and the latter is highly affected by the value and the evolution of the biomarker over time. The SLD

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Pharmaceutical Statistics* published by John Wiley & Sons Ltd.

distribution is often characterized by an excess of zeros and right skewness. Patients whose treatment removes all visible signs of the disease generates zero values for the SLD. This excess of zeros is therefore highly informative of treatment efficiency.

In previous work, including ours, a conditional two-part joint model (C-TPJM) has been introduced to fit a longitudinal biomarker jointly with the risk of terminal event, while taking into account the semicontinuous distribution of the biomarker.^{1,2} When an excess of true zeros is observed, the model was shown superior to standard approaches such as left-censoring the biomarker's distribution (i.e., assuming zero values are censored values, too small to be observed) to compare clinical treatment strategies. Indeed, the left-censoring one-part joint model (OPJM) fails to explain some informative variability in the biomarker evolution over time because true zeros (i.e., not censored) are observed, resulting in a reduced discrimination of the risk of death between treatment arms. The conditional two-part model³ decomposes the distribution of the outcome into a binary part corresponding to zero versus positive values and a continuous part with positive values only, both outcomes being modeled by a mixed effects regression model. The binary and continuous parts are linked through correlated random effects. The model yields covariate effects, such as treatment effect, on the probability of observing a positive versus zero SLD in the binary part and on the expected value of the biomarker conditional on observing a positive value (i.e., zeros excluded) in the continuous part. On the other hand, this model cannot provide treatment effect on the marginal mean of the biomarker, which is often of clinical interest. For the terminal event, the hazard function can be expressed conditionally on observing a zero SLD value (which is indicative of a complete response to treatment) and on the expected value of the biomarker among positive values (which is indicative of a partial response to treatment). This type of association structure could help in the evaluation of treatment effect on survival through its indirect effect on the biomarker.

Outside the framework of joint modeling, a marginalized two-part model^{4,5} has been proposed as an alternative to the conditional two-part model, useful when the interest lies in the population-average effects of covariates, such as treatment effect, on the biomarker. This model accounts for the zero values in the continuous part of the model and provides covariate effects on the marginal mean of the biomarker. In addition, a binary part, similar to the conditional two-part model, accounts for the excess of zeros and can assess covariate effects on the probability of observing a positive biomarker value vs. a zero value. The conditional and marginalized two-part models can address different clinical questions. When the interest is in the expectation of the biomarker among positive values, the conditional model is more appropriate while the marginalized two-part model may lead to arbitrary heterogeneity and provides less interpretable estimates on the conditional mean of the biomarker among positive values.⁵ However, in many clinical applications such as cancer clinical trials, the interest is often to assess covariates effect on the marginal mean of a biomarker (i.e., the effect on the whole population of interest), in particular to obtain FDA approvals, rather than on the mean conditional on a positive value (i.e., a subpopulation not necessarily representative of the whole population). The left-censoring one-part model provides similar covariate effects on the marginal mean of the biomarker as the marginalized two-part model, but does not account for the excess of zero values. The left-censoring OPJM rather considers an excess of values under a certain threshold or limit of detection. The marginalized two-part model combines the advantages of the conditional two-part model and the left-censoring one-part model by allowing a direct interpretation of covariate effect on the population mean value of the biomarker while also accounting for the excess of zeros. Shahrokhbabadi et al.⁶ recently proposed a marginalized two-part joint model for the joint analysis of medical costs and death, however the approach is limited in complexity (only random intercepts, constant baseline risk, NLMIXED estimation procedure) and it is not properly contrasted with the existing literature (i.e., one-part joint and conditional two-part joint models). The model has indeed a structure similar to that of the conditional two-part joint model proposed by Liu,¹ where the two components of the two-part model (binary and continuous) are linked through a shared subject-specific random intercept while another subject-specific random intercept captures the residual individual variability in the continuous part. Both random effects are shared with a Cox proportional hazards model for a terminal event. The interpretation of the random intercept in the binary part of the GLMM is complicated since it captures both the correlation among the repeated measures over time and the correlation among the two components of the semicontinuous model.

In this article, we introduce an alternative and more flexible formulation of the M-TPJM. It accounts for the relationship between the binary and continuous parts through correlated random effects. Moreover, the estimation is performed through a penalized likelihood approach (allowing for a smooth and flexible non-parametric baseline hazard curve) and includes a Monte-Carlo numerical approximation for the integration over the multivariate Gaussian random effects, leading to better scaling properties than the Gauss-Hermite strategy of Shahrokhbabadi et al.⁶ Indeed, it avoids the curse of dimensionality issue and thus allows more complex models to be fitted (e.g., including random slopes). We also contrast the marginalized formulation of the two-part joint model with both the conditional formulation and the

one part joint model, showing how this marginalized model combines the properties of the two latter models for better clinical decision making.⁷ Finally, we introduce two different association structures between the semicontinuous biomarker and the risk of event, namely the “shared random effects” and the “current value” parameterizations.

In the application section of this article, we illustrate the differences between the three modeling strategies for a semicontinuous biomarker and a terminal event (i.e., left-censoring OPJM, C-TPJM, and M-TPJM). Besides, we illustrate how the M-TPJM facilitates the decision-making for clinicians by providing the effect of covariates such as treatment on the marginal mean of the biomarker and its subsequent effect on the risk of death.

In this article, we propose a marginalized two-part joint model (M-TPJM) for a longitudinal semicontinuous outcome and a terminal event. We compare the M-TPJM with the C-TPJM and the left-censoring OPJM through simulation studies and provide a detailed interpretation of these models. The remainder of the article is structured as follows: in Section 2, we describe the M-TPJM and its estimation method. In Section 3, we present a simulation study to assess the performance of the model and compare it to competing approaches that treat the excess of zeros differently. An application to a randomized clinical trial comparing a combination of chemotherapy and panitumumab (anti-EGFR monoclonal antibody) to chemotherapy alone, in patients with metastatic and/or recurrent squamous-cell carcinoma of the head and neck, is proposed in Section 4 and we conclude with a discussion in Section 5.

2 | MODEL

2.1 | Left-censoring one-part model for the biomarker

Let Y_{ij} denote the biomarker value for subject i ($i = 1, \dots, n$), at visit j ($j = 1, \dots, n_i$). The model assumes the biomarker can be subject to left-censoring when it decreases below a limit of detection c .

$$Y_{ij}^* = \begin{cases} Y_{ij} & \text{if } Y_{ij} > c \\ 0 & \text{otherwise} \end{cases}$$

The Y_{ij}^* has the same distribution as the Y_{ij} when $Y_{ij} > c$. For the observations $Y_{ij} = 0$, all we know is $P(Y_{ij}^* = 0) = P(Y_{ij} < c)$, see Tobin.⁸ The density function is used for the non-censored biomarker values while the cumulative distribution function is used for the censored observations (the censoring threshold is chosen as the smallest positive value observed in the data when not provided by the investigators). Below we assume a log-normal distribution for the biomarker.

2.2 | Conditional two-part model for the biomarker

The biomarker distribution is decomposed into a binary outcome $I[Y_{ij} > 0]$ and a positive-continuous outcome $Y_{ij} | Y_{ij} > 0$. A GLMM with a logit link is assumed for the binary outcome and a log-normal mixed effects regression model for the continuous outcome. The logarithm link in the continuous part is used to linearize the biomarker evolution over time and correct for right-skewness. The two components are linked through correlated random effects. The two-part model for the biomarker is defined as follows:

$$\begin{cases} \text{logit}(\text{Prob}(Y_{ij} > 0 | \mathbf{a}_i^C)) = \mathbf{X}_{Aij}^T \boldsymbol{\alpha}^C + \mathbf{Z}_{Aij}^T \mathbf{a}_i^C & \text{(Binary part),} \\ \log(E[Y_{ij} | Y_{ij} > 0, \mathbf{b}_i^C]) = \mathbf{X}_{Bij}^T \boldsymbol{\beta}^C + \mathbf{Z}_{Bij}^T \mathbf{b}_i^C & \text{(Continuous part),} \end{cases}$$

where \mathbf{X}_{Aij} and \mathbf{Z}_{Aij} are vectors of covariates associated with the fixed effect parameters $\boldsymbol{\alpha}^C$ and the random effects \mathbf{a}_i^C for the binary part. Similarly, \mathbf{X}_{Bij} and \mathbf{Z}_{Bij} are vectors of covariates associated with the fixed effect parameters $\boldsymbol{\beta}^C$ and the random effects \mathbf{b}_i^C . We assume a log-normal distribution for $Y_{ij} | Y_{ij} > 0$ in the continuous part with location parameter given in Section 2.7 (Equation 4) and error $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^C)$. The two vectors of random effects follow a multivariate normal distribution:

$$\begin{bmatrix} \mathbf{a}_i^C \\ \mathbf{b}_i^C \end{bmatrix} \sim MVN \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_a^2 & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b^2 \end{bmatrix} \right), \text{ where } a \text{ and } b \text{ can be multidimensional.}$$

The vectors of correlated subject-specific random effects \mathbf{a}_i^C and \mathbf{b}_i^C account for the correlation between repeated measurements within an individual and the correlation between the two components of the model. The logistic regression model includes covariates that represent the effect of an individual's characteristics on the probability of observing a positive versus zero biomarker value. The continuous part represents the log of the expected value of the biomarker given a positive biomarker value. This model differs from the conditional two-part model proposed in Rustand et al.² for which the continuous part represented the expected value of the log-transformed longitudinal outcome, resulting in an additive effect of covariates on the transformed scale of the biomarker. Using now a logarithm link facilitates the interpretation of a covariate k , where $\exp(\beta_k)$ represents the multiplicative effect on the (original scale) biomarker value at a given time point, conditional on a positive value at that time point, associated with a one-unit increase in the covariate X_k .⁹ To illustrate, let us consider our application to tumor size. The formulation presented in Rustand et al.² could indicate that treatment X reduces tumor size by Y units on the logarithmic scale per unit of time. In contrast, our newly proposed formulation suggests that the treatment reduces tumor size by Z% on the original scale per unit of time, which is clinically more appealing. Additionally, we can measure uncertainty and significance of covariate effects directly on this scale unlike our previous approach. It should also be noted that in Rustand et al.,² the exponential of the covariate effects can be interpreted in terms of change in the median of the response variable, which can also be of scientific interest. We assume the risk of terminal event depends on the repeated measurements of the biomarker, with two different association structures: shared random effects and current level of the biomarker (See Section 2.5).

2.3 | Marginalized two-part model for the biomarker

In the context of two-part models, the term “marginalized” refers to the biomarker distribution including both zeros and positive values. It will be compared with the “conditional form” which includes a positive-continuous outcome (i.e., constraint model upon positive values of the outcome) in addition to the binary outcome. The conditional form does not refer here to the “subject-specific” usage, that is, conditional on random effects. In this article, we will use the terminology “population average/subject-specific” to avoid some confusion with the “marginalized/conditional” concept.

In the M-TPJM, the binary part is similar to the one used in the conditional model, but the continuous part models the covariate effects on the marginal mean of the biomarker. The model is defined as follows:

$$\begin{cases} \text{logit}(\text{Prob}(Y_{ij} > 0 | \mathbf{a}_i^M)) = \mathbf{X}_{Aij}^\top \boldsymbol{\alpha}^M + \mathbf{Z}_{Aij}^\top \mathbf{a}_i^M & \text{(Binary part),} \\ \log(E[Y_{ij} | \mathbf{b}_i^M]) = \mathbf{X}_{Bij}^\top \boldsymbol{\beta}^M + \mathbf{Z}_{Bij}^\top \mathbf{b}_i^M & \text{(Continuous part),} \end{cases} \quad (1)$$

The marginalized two-part model gives the effect of covariates on the marginal mean of the biomarker instead of the mean conditional on observing a positive value of the biomarker by including both the zeros and positive values in the continuous part. We assume a log-normal distribution for Y_{ij} in the continuous part with location parameter given in Section 2.7 (Equation 3) and error $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^M)$. The random effects \mathbf{a}_i^M and \mathbf{b}_i^M follow a multivariate normal distribution. They capture some correlation due to potentially unobserved process driving the probability of positive value and the marginal mean value, that is, lower values of the biomarker are more likely correlated with the probability of observing a zero. Another induced correlation is that the expression of the overall mean also depends on the probability of observing a positive value (see Equation 3). With the conditional two-part model, the association between the binary and continuous part is only captured through the correlation structure of the random effects.

2.4 | Survival component of the joint model

A Cox proportional hazards model is used for the terminal event.

$$\lambda_i(t|\cdot) = \lambda_0(t) \exp\left(\mathbf{X}_{Si}^\top \boldsymbol{\gamma} + \mathbf{h}(\cdot)^\top \boldsymbol{\varphi}\right)$$

The effect of covariates \mathbf{X}_{Si} on the risk of death is given by the vector of regression coefficients $\boldsymbol{\gamma}$. The function $\mathbf{h}(\cdot)$ corresponds to the association function between the biomarker and the risk of event, and $\boldsymbol{\varphi}$ the corresponding vector of association parameters (i.e., fixed effects that scale the association term). The association function corresponds to the elements that are shared to define the association between the biomarker and the survival, it is a function of the fixed and random effects from the biomarkers' mixed effects models.

2.5 | Association structures

We propose two possible association structures. In the “shared random effects” (SRE) association structure between the M-TPJM and the survival model, we have

$$\lambda_i(t|\mathbf{a}_i^M, \mathbf{b}_i^M) = \lambda_0(t) \exp(\mathbf{X}_{Si}^\top \boldsymbol{\gamma}^M + \mathbf{a}_i^{M\top} \boldsymbol{\varphi}_a^M + \mathbf{b}_i^{M\top} \boldsymbol{\varphi}_b^M),$$

and in the “current level” (CL) association (also referred to as “current value”) we have

$$\lambda_i(t|\mathbf{b}_i^M) = \lambda_0(t) \exp(\mathbf{X}_{Si}^\top \boldsymbol{\gamma}^M + E[Y_i(t)|\mathbf{b}_i^M] \boldsymbol{\varphi}^M),$$

where $Y_i(t)$ is the expected value of the longitudinal marker for individual i at time t , given by the two-part submodel. The SRE association is useful to explore the association between an individual's deviation from the population mean evolution of the biomarker and the risk of terminal event but does not take into account the covariance between the two vectors of random effects in the survival model. The difference in the SRE association structure between the M-TPJM and the C-TPJM is that the individual heterogeneity captured in the continuous part by the random effects is conditional on observing a positive value of the biomarker with the C-TPJM, while for the marginalized model it corresponds to the entire population. The biomarker model takes into account informative censoring by the terminal event through the shared random effects while the survival part gives the hazard ratio of the covariates conditional on the random effects, assuming proportional hazards.

In the CL association, $\boldsymbol{\varphi}$ is a scalar parameter that quantifies the strength of the association between the true unobserved value of the longitudinal biomarker (that is the error-free value of the biomarker) and the risk of event. With a C-TPJM, the “current value” association with the biomarker is defined as:

$$\begin{aligned} \lambda_i(t|\mathbf{a}_i^C, \mathbf{b}_i^C) &= \lambda_0(t) \exp(\mathbf{X}_{Si}^\top \boldsymbol{\gamma}^C + E[Y_i(t)|\mathbf{a}_i^C, \mathbf{b}_i^C] \boldsymbol{\varphi}^C), \\ \text{with } E[Y_i(t)|\mathbf{a}_i^C, \mathbf{b}_i^C] &= \text{Prob}(Y_i(t) > 0|\mathbf{a}_i^C) E[Y_i(t)|Y_i(t) > 0, \mathbf{b}_i^C], \end{aligned} \quad (2)$$

that is a combination of two non-linear processes, imposing a specific functional form for the association of the biomarker current value with the risk of event. The M-TPJM directly models the mean value of the biomarker $E[Y_i(t)|\mathbf{b}_i^M]$, which facilitates the interpretation of covariates effect on the biomarker mean value. Because $E[Y_i(t)|\mathbf{b}_i^M]$ is directly obtained from the M-TPJM, the variance of the estimated value of the biomarker is reduced under the M-TPJM compared to the C-TPJM, as illustrated in Figure 2. In terms of covariate effects, under the CL association, a covariate can be associated with the risk of event through its own association with the biomarker or independently of the biomarker when it is also included in the survival model.

2.6 | Interpretation of treatment effect

It is simple to estimate both subject-specific and population average means of the biomarker under the M-TPJM (and of the biomarker conditional on a positive value under the C-TPJM), particularly for covariates not included as random effects as the regression coefficients takes both subject-specific and population average interpretations (see appendix B of Smith¹⁰).

For the M-TPJM, the population average mean of the biomarker is expressed as

$$E[Y_{ij}] = E_{\mathbf{b}_i^M} [E(Y_{ij}|\mathbf{b}_i^M)] = \exp\left(\mathbf{X}_{Bij}^\top \boldsymbol{\beta}^M + \frac{1}{2} \mathbf{Z}_{Bij}^\top \boldsymbol{\Sigma}_{bb}^M \mathbf{Z}_{Bij}\right)$$

While for the C-TPJM, the population average mean of the biomarker conditional on a positive value is expressed as:

$$E[Y_{ij}|Y_{ij} > 0] = E_{\mathbf{b}_i^C} [E(Y_{ij}|Y_{ij} > 0, \mathbf{b}_i^C)] = \exp\left(\mathbf{X}_{Bij}^\top \boldsymbol{\beta}^C + \frac{1}{2} \mathbf{Z}_{Bij}^\top \boldsymbol{\Sigma}_{bb}^C \mathbf{Z}_{Bij}\right),$$

where $\boldsymbol{\Sigma}_{bb}^M$ and $\boldsymbol{\Sigma}_{bb}^C$ corresponds to the variance–covariance of random effects from the continuous part with the M-TPJM and the C-TPJM, respectively.

From these expressions, it is clear that $\exp(\beta_{trt}^M)$ with the M-TPJM (respectively $\exp(\beta_{trt}^C)$ for the C-TPJM), where β_{trt} is the treatment effect in the continuous part of the model, corresponds to the multiplicative effect on both the subject-specific and the population average value of the biomarker (respectively the value of the biomarker conditional on observing a positive value with the C-TPJM), when treatment is not included as a random effect. We can get an approximation of the population average effect of treatment on the biomarker with the C-TPJM, but this effect is conditional on the random effects and the value of other covariates included in the model. Moreover, the delta method or resampling techniques must be employed to get a confidence interval and a Wald test on this population average effect of treatment.

A diagram giving an interpretation of the treatment effect under the C-TPJM and the M-TPJM is displayed in Figure 1. With the SRE association, the survival model gives the hazard ratio of treated vs. untreated patients and the

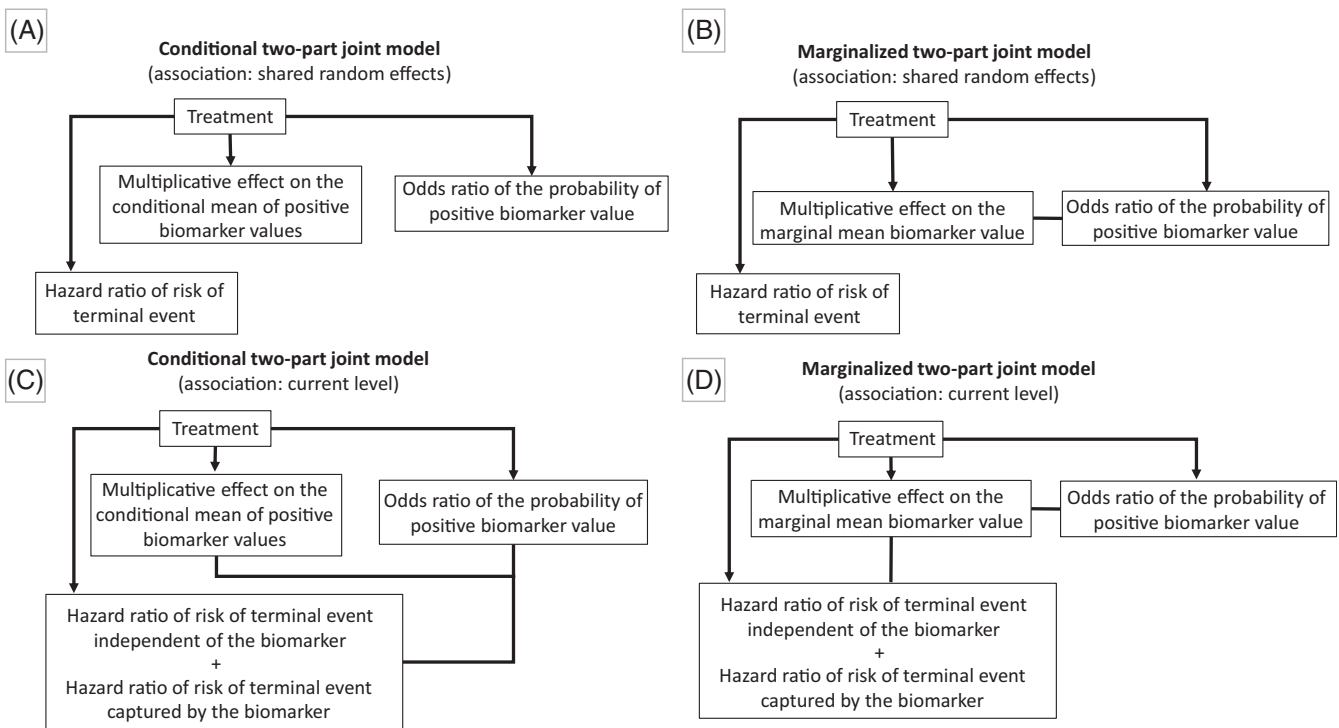


FIGURE 1 Diagrams describing the interpretation of the treatment effect with the C-TPJM (left) and the M-TPJM (right) for the shared random effects (up) and the current level (down) association structures. The marginalized model includes zero values to give the effect of treatment on the marginal mean biomarker value. The shared random effects association provides a hazard ratio for treatment effect on the risk of event adjusted for unmeasured confounders captured by the random effects. The current level association structure accounts for a treatment effect on the risk of event independent on the biomarker as well as a possible treatment effect captured by the time-dependent association with the biomarker on the risk of event.

biomarker model gives the effect of treatment on the probability of observing a positive value of the biomarker in the binary part.

With the CL association, the treatment effect in the biomarker model affects also the survival part through the association between the biomarker and the risk of event. This results in the treatment effect to vary over time in the biomarker model and in a non-proportional effect on the survival model (i.e., the hazards are proportional conditional on the time-dependent current level of the biomarker). We recommend to use graphical representations to get a clear idea of the time-dependent effect of treatment on survival time with the CL association structure as illustrated in Figure S1 of the supplementary material for the models from the application Section.

We can compute the subject-specific (i.e., conditional on the random effects) time-dependent overall treatment effect with the CL association. It corresponds to the treatment effect for the average patient, with random effects equal to zero. Moreover, it is possible to compute the average treatment effect in the population from the subject-specific one using Monte-Carlo simulations, as discussed in van Oudenhoven et al.¹¹

2.7 | Estimation procedure

The full likelihood of the M-TPJM is given by

$$L(\cdot) = \int_{\mathbf{a}_i^M} \int_{\mathbf{b}_i^M} \prod_{j=1}^{n_i} \text{Prob}(Y_{ij} > 0 | \mathbf{a}_i^M)^{U_{ij}} (1 - \text{Prob}(Y_{ij} > 0 | \mathbf{a}_i^M))^{(1-U_{ij})} \\ \times \left\{ \frac{1}{(\sqrt{2\pi\sigma_\epsilon^M})} Y_{ij}^{-1} \exp\left(-\frac{(\log(Y_{ij}) - \mu_{ij}^M)^2}{2\sigma_\epsilon^M}\right) \right\}^{U_{ij}} \\ \times \lambda_i(T_i | \Theta)^{\delta_i} \exp\left(-\int_0^{T_i} \lambda_i(t | \Theta) dt\right) p(\mathbf{a}_i^M, \mathbf{b}_i^M) d\mathbf{b}_i^M d\mathbf{a}_i^M$$

With $U_{ij} = I[Y_{ij} > 0]$, $\Theta = (\boldsymbol{\alpha}^M, \boldsymbol{\beta}^M, \mathbf{a}_i^M, \mathbf{b}_i^M, \boldsymbol{\gamma}^M, \boldsymbol{\varphi}^M)$ and μ_{ij}^M the location parameter of the log-normal distribution $\log\mathcal{N}(\mu_{ij}^M, \sigma_\epsilon^M)$ for the contribution of positive biomarker values to the likelihood. Details on the construction of the log-likelihood is given in the supplementary material.

With a M-TPJM, the marginal mean of Y_{ij} is:

$$E[Y_{ij} | \mathbf{b}_i^M] = \text{Prob}(Y_{ij} > 0 | \mathbf{a}_i^M) \exp\left(\mu_{ij}^M + \frac{\sigma_\epsilon^M}{2}\right)$$

Using the parameterization from Equation (1), we can derive the corresponding location parameter of the log-normal distribution as:

$$\mu_{ij}^M = \mathbf{X}_{Bij}^\top \boldsymbol{\beta}^M + \mathbf{Z}_{Bij}^\top \mathbf{b}_i^M - \log(\text{Prob}(Y_{ij} > 0 | \mathbf{a}_i^M)) - \frac{\sigma_\epsilon^M}{2} \quad (3)$$

With a C-TPJM, the likelihood contributions from the binary part and the continuous part are only linked through the random effects correlation structure. The location parameter of the log-normal distribution for the positive values is therefore,

$$\mu_{ij}^C = \mathbf{X}_{Bij}^\top \boldsymbol{\beta}^C + \mathbf{Z}_{Bij}^\top \mathbf{b}_i^C - \frac{\sigma_\epsilon^C}{2}, \quad (4)$$

The baseline hazard in the survival part of the model is approximated with m cubic M-splines with Q knots. A penalization term ensure that the baseline hazard is smooth,

$$pl(\Theta) = l(\Theta) - \kappa \int_0^{\infty} \lambda_0''(t)^2 dt,$$

where $l(\Theta) = \sum_{i=1}^n \log(L_i(\Theta))$ and κ a smoothing parameter chosen using an approximate cross-validation criterion from a separate Cox model.

We use the Levenberg–Marquardt algorithm¹² to maximize the log-likelihood. The integration over the random effects has no analytical solution, therefore we approximate their value with a Monte-Carlo method. The number of points for the Monte-Carlo integration methods defines the tradeoff between the precision of the approximation of the random effects distribution and the computation time.

The approximated likelihood cross-validation (LCV) criterion¹³ for evaluation of the goodness-of-fit of the models can be used as a model choice criterion. It corresponds to the Akaike information criterion (AIC) in the case of the penalized maximum likelihood estimation. The LCV requires the outcome to be the same and while the overall mean is the same between the left-censoring OPJM and the M-TPJM, the latter does include the contribution of the binary component into the likelihood, making them not comparable according to this criterion. However, the LCV can compare the goodness-of-fit between the C-TPJM and the M-TPJM as well as the SRE and CL association structures for each type of joint model. The left-censoring OPJM, the C-TPJM and the M-TPJM are estimated with the function *longiPenal* of the R package *frailtypack*, available on the comprehensive R archive network (CRAN).

3 | SIMULATION STUDY

3.1 | Simulation study design

We conducted simulation studies to compare the left-censoring OPJM, the conditional TPJM and the marginalized TPJM in terms of bias and coverage probabilities. We propose three scenarios where the true model for data generation is either the M-TPJM (scenario 1), the C-TPJM (scenario 2) or the left-censoring OPJM (scenario 3). The parameters used for the data generation are based on the results from the real data application, resulting in about 10% zero values in the biomarker distribution. An additional set of simulations is presented in the supplementary material with an increased zero rate (20%). For each scenario, 300 datasets are generated with 400 individuals each. The number of datasets was chosen to obtain a minimal Monte Carlo standard error of the parameter estimates under a reasonable computation time (about 4 days per scenario with MPI parallelization over 80 CPUs). We focus on the CL association structure for these simulations since it is a more challenging joint model to estimate (the survival model requires an additional integration step in the optimization procedure). Moreover, the CL association structure provides a slightly better fit compared to the SRE association structure in our application. For the data generation assuming the M-TPJM as the true model, we first generate the zero values from a Bernoulli distribution, then the longitudinal biomarker measurements assuming a log-normal distribution for the positive biomarker values, using the location parameter of Equation (3). The longitudinal measurements are generated for the entire follow-up and then we use the R package *PerMAlgo* to generate random death times that depends on the time-dependent biomarker value and random censoring times.¹⁴ The data generation for the C-TPJM is similar, except that the location parameter does not include the linear predictor from the binary part (Equation 4). The observed value of the biomarker with the C-TPJM is therefore defined by Equation (2), which is non-linear on the log scale. For the one-part model, we generate the longitudinal measurements and then the zero excess with a censoring threshold chosen as the first (or second) decile of the distribution. The number of repeated measurements of the biomarker per individual varies between 1 and 16, with a median of 2. The percentage of patients who die during the 4 years follow-up is 80% following the real data death rate. Therefore, most of the biomarker observations are in the early follow-up (the sample size decreases over time as censoring and death occurs). A binary covariate generated from a Bernoulli distribution with $p = 0.5$ corresponding to the treatment effect is included in each submodel of the joint model, with a time-interaction for each submodel of the two-part models.

We use the same parameters as in the application to choose the number of knots for the splines baseline hazard approximation (5 knots). The penalization term was chosen by cross-validation from an univariate survival model. We use 1000 Monte-Carlo integration points for the numerical approximation of the integral over the random effects distribution. A replication script is available at github.com/DenisRustand/TPJM_sim; it includes the code “MTPJM_sim.R” for the generation and estimation of a M-TPJM and the code “TPJM_sim.R” for the generation and estimation of a C-TPJM and a OPJM.

The parameters of the binary part can be compared between the C-TPJM and the M-TPJM as they both give the effect of covariates on the probability to observe a positive value. The parameters of the continuous part can be compared between the left-censoring OPJM and the M-TPJM as they both give the effect of covariates on the marginal mean of the biomarker. The continuous part of the C-TPJM cannot be directly compared to the continuous part of the other two models. The direct effect of treatment on the risk of death and the association between the biomarker and the risk of death in the survival part can be compared between the three models.

3.2 | Results

Results from the simulation study are presented in Tables 1–3 and Tables S1–S3 of the supporting information.

3.2.1 | Scenario 1 – True model: M-TPJM

The M-TPJM recovers the true parameters value with good accuracy, coverage probabilities are close to 95% (Table 1). Fixed effects parameters for the continuous part of the left-censoring OPJM are biased, with an intercept value $\hat{\beta}_0 =$

TABLE 1 Summary of the results of simulations scenario 1 (true model: marginalized TPJM), 300 datasets with 400 individuals each and 1000 integration points, 10.17% zeros on average (SD = 1.33). The true value of the parameters estimated in the continuous part of the C-TPJM are unknown, therefore coverage probabilities are not provided for these parameters.

Variable	Left-censoring OPJM	C-TPJM	M-TPJM	
	Est. ^a (SD) ^b [CP] ^c	Est. (SD) [CP]	Est. (SD) [CP]	
Binary part				
Intercept	$\alpha_0 = 6$	6.09 (0.64) [96%]	6.11 (0.57) [94%]	
Time	$\alpha_1 = -3$	-3.04 (0.44) [96%]	-3.04 (0.35) [93%]	
Treatment	$\alpha_2 = 1$	0.93 (0.75) [96%]	0.94 (0.68) [96%]	
Time: treatment	$\alpha_3 = -2$	-1.93 (0.65) [94%]	-1.95 (0.53) [94%]	
Continuous part				
Intercept	$\beta_0 = 1.5$	1.69 (0.06) [06%]	1.52 (0.05)	1.53 (0.05) [90%]
Time	$\beta_1 = -0.5$	-0.58 (0.10) [93%]	-0.35 (0.03)	-0.50 (0.06) [93%]
Treatment	$\beta_2 = 0.3$	0.38 (0.08) [77%]	0.28 (0.07)	0.30 (0.07) [93%]
Time: treatment	$\beta_3 = 0.3$	-0.13 (0.17) [22%]	0.42 (0.09)	0.30 (0.08) [95%]
Residual S.E.	$\sigma_\varepsilon = 0.3$	0.64 (0.07) [00%]	0.32 (0.01)	0.30 (0.01) [92%]
Survival part				
Treatment	$\gamma = -0.2$	-0.16 (0.13) [92%]	-0.16 (0.12) [91%]	-0.18 (0.12) [92%]
Association	$\varphi = 0.08$	0.09 (0.02) [94%]	0.08 (0.02) [95%]	0.08 (0.02) [95%]
Random effects				
Intercept (binary part)	$\sigma_a = 1.4$		1.33 (0.28)	1.37 (0.28)
Intercept (continuous part)	$\sigma_{b_0} = 0.6$	0.45 (0.06)	0.62 (0.03)	0.61 (0.03)
Slope (continuous part)	$\sigma_{b_1} = 0.3$	0.69 (0.12)	0.33 (0.07)	0.28 (0.08)
	$cor_{ab_0} = 0.5$		0.51 (0.17)	0.56 (0.16)
	$cor_{ab_1} = 0.5$		0.07 (0.30)	0.45 (0.30)
	$cor_{b_0b_1} = 0.2$	0.20 (0.23)	-0.20 (0.17)	0.27 (0.24)
Convergence rate		100%	100%	100%

^aMean of parameter estimates.

^bStandard deviation from the mean.

^cCoverage probability.

1.69 (SD = 0.06, CP = 6%) where the true value is $\beta_0 = 1.5$. The model is not able to handle properly the excess of zero values. We illustrate this systematic bias with a plot of the estimated mean trajectory of the biomarker compared to the true trajectory under the three scenarios (Figure 2). The time by treatment interaction effect under the left-censoring OPJM is negative ($\hat{\beta}_3 = -0.13$, SD = 0.17, CP = 22%) where the true value is positive ($\beta_3 = 0.3$). The binary part of the C-TPJM gives unbiased results. For the survival part, both the left-censoring OPJM and the C-TPJM are able to capture the treatment effect on the risk of death independent of the biomarker ($\hat{\gamma} = -0.16$, SD = 0.13, CP = 92% for the left-censoring OPJM and $\hat{\gamma} = -0.16$, SD = 0.12, CP = 91% for the C-TPJM), with a mean value slightly lower than the true value ($\gamma = -0.2$) compared to the M-TPJM ($\hat{\gamma} = -0.18$, SD = 0.12, CP = 92%). The association between the biomarker and the survival is also unbiased for the left-censoring OPJM ($\hat{\varphi} = 0.09$, SD = 0.02, CP = 94%) and the C-TPJM ($\hat{\varphi} = 0.08$, SD = 0.02, CP = 95%) where the true value is $\varphi = 0.08$. The standard deviations of the random effects are properly estimated with the two TPJMs but not with the left-censoring OPJM, the random intercept ($\sigma_{b_0} = 0.6$) is biased downwards ($\hat{\sigma}_{b_0} = 0.45$, SD = 0.06) and the random slope ($\sigma_{b_1} = 0.3$) is biased upwards ($\hat{\sigma}_{b_1} = 0.69$, SD = 0.12). The correlation between the random intercept and slope in the continuous part ($corr_{b_0b_1} = 0.2$) is biased with the C-TPJM ($\hat{corr}_{b_0b_1} = -0.20$, SD = 0.17) as well as the correlation between the intercept from the binary part and the slope in the continuous part ($corr_{ab_1} = 0.5$), finding no correlation ($\hat{corr}_{ab_1} = 0.07$, SD = 0.30).

3.2.2 | Scenario 2 – True model: C-TPJM

The parameter estimates in the binary part ($\alpha_0 = 6, \alpha_1 = -3, \alpha_2 = 1, \alpha_3 = -2$) are biased with the M-TPJM ($\hat{\alpha}_0 = 5.46$, SD = 0.57, CP = 69%; $\hat{\alpha}_1 = -2.34$, SD = 0.38, CP = 39%; $\hat{\alpha}_2 = 0.66$, SD = 0.74, CP = 89%; $\hat{\alpha}_3 = -1.45$, SD = 0.62, CP = 69%) while the C-TPJM is unbiased with similar variability (Table 2). This could be due to the correlation between the binary part and the continuous part in the M-TPJM (Equation 3), while they are simulated independent conditional on the random effects. As displayed in Figure 2, the mean behavior of the biomarker is not linear on the log scale with the C-TPJM as opposed to the left-censoring OPJM and the M-TPJM. In particular, the mean value of the biomarker converges towards zero at the end of the follow-up because the probability of positive value decreases over time in the binary part. The M-TPJM is not able to capture this trend in the late follow-up (i.e., where there are less observations because some patients got censored or died during follow-up). Although the biomarker average trajectory specified as log-linear for the M-TPJM does not fit well, it can be improved by including more general functions of time in the linear predictor of the continuous part. This is illustrated by including splines in the linear predictor of the continuous part of the M-TPJM to capture a non-linear trend of the population average biomarker trajectory, see Figure S2 of the supplementary material for an illustration. The left-censoring OPJM seems severely biased for this simulation scenario,

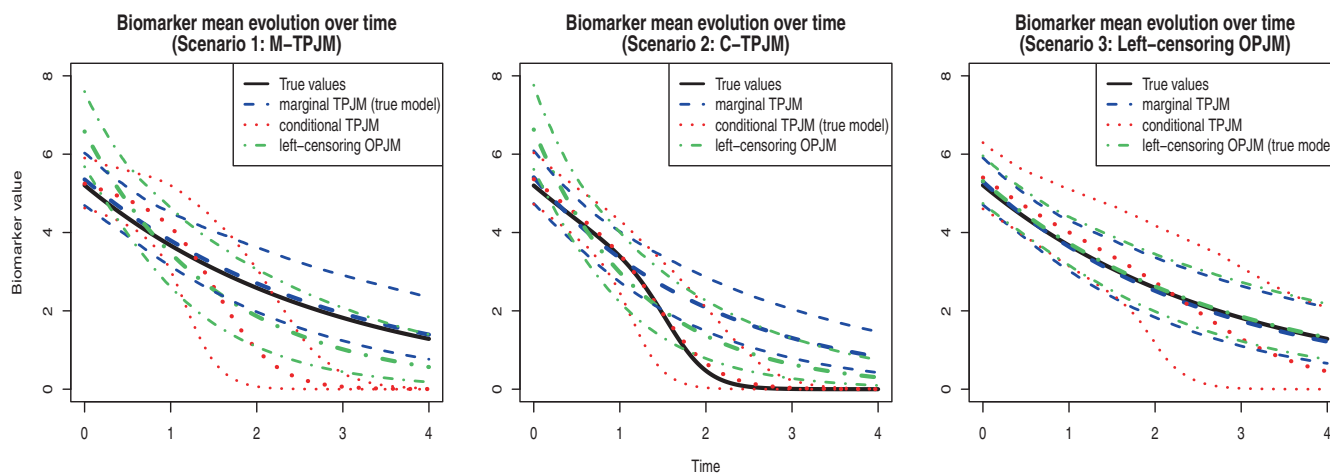


FIGURE 2 Mean biomarker trajectory captured in the simulation studies (Tables 1–3) for the M-TPJM, the C-TPJM and the left-censoring OPJM compared to the true trajectory. The C-TPJM always assume a non-linear population average biomarker trend as it is defined by the combination of the binary and continuous parts while the OPJM and M-TPJM can define the population average biomarker trajectory as log-linear (although the trajectory can be more flexible when including functions of time in the linear predictor of the continuous part).

TABLE 2 Summary of the results of simulations scenario 2 (true model: conditional TPJM), 300 datasets with 400 individuals each and 1000 integration points, 10.53% zeros on average (SD = 1.36). The true value of the parameters estimated in the continuous part of the left-censoring OPJM and the M-TPJM are unknown, therefore coverage probabilities are not provided for these parameters.

Variable	Left-censoring OPJM			C-TPJM	M-TPJM
	Est. ^a (SD ^b) [CP ^c]			Est. (SD) [CP]	Est. (SD) [CP]
Binary part					
Intercept	$\alpha_0 = 6$			6.13 (0.64) [96%]	5.46 (0.57) [69%]
Time	$\alpha_1 = -3$			-3.07 (0.45) [95%]	-2.34 (0.38) [39%]
Treatment	$\alpha_2 = 1$			1.03 (0.85) [96%]	0.66 (0.74) [89%]
Time: treatment	$\alpha_3 = -2$			-2.04 (0.72) [95%]	-1.45 (0.62) [69%]
Continuous part					
Intercept	$\beta_0 = 1.5$	1.68 (0.07)		1.53 (0.05) [90%]	1.53 (0.05)
Time	$\beta_1 = -0.5$	-0.68 (0.10)		-0.50 (0.06) [90%]	-0.60 (0.06)
Treatment	$\beta_2 = 0.3$	0.41 (0.08)		0.30 (0.07) [91%]	0.32 (0.07)
Time: treatment	$\beta_3 = 0.3$	-0.24 (0.16)		0.30 (0.08) [94%]	0.23 (0.09)
Residual S.E.	$\sigma_\varepsilon = 0.3$	0.63 (0.08)		0.30 (0.01) [94%]	0.30 (0.01)
Survival part					
Treatment	$\gamma = -0.2$	-0.21 (0.13) [95%]		-0.20 (0.12) [95%]	-0.21 (0.12) [95%]
Association	$\varphi = 0.08$	0.10 (0.03) [92%]		0.08 (0.02) [94%]	0.08 (0.02) [93%]
Random effects					
Intercept (binary part)	$\sigma_a = 1.4$			1.35 (0.29)	1.25 (0.28)
Intercept (continuous part)	$\sigma_{b_0} = 0.6$	0.47 (0.07)		0.61 (0.03)	0.61 (0.03)
Slope (continuous part)	$\sigma_{b_1} = 0.3$	0.80 (0.16)		0.29 (0.05)	0.37 (0.06)
	$cor_{ab_0} = 0.5$			0.53 (0.16)	0.54 (0.17)
	$cor_{ab_1} = 0.5$			0.51 (0.25)	0.62 (0.20)
	$cor_{b_0b_1} = 0.2$	0.18 (0.19)		0.20 (0.19)	0.34 (0.18)
Convergence rate		100%		100%	100%

^aMean of parameter estimates.

^bStandard deviation from the mean.

^cCoverage probability.

especially for the time by treatment interaction effect on the marginal mean of the biomarker ($\hat{\beta}_3 = -0.24$, SD = 0.16). As observed in the first scenario, the effect of treatment on the risk of event independent of the biomarker and the association parameter are properly recovered for the three models except the left-censoring OPJM with a slightly higher estimate and standard error for the association ($\hat{\varphi} = 0.10$, SD = 0.03) than the true value recovered by the M-TPJM and the C-TPJM ($\hat{\varphi} = 0.08$, SD = 0.02) while coverage probabilities are close to 95% with the three models. The standard deviations of the random effects and their correlation is properly captured with the C-TPJM and similarly with the M-TPJM while the left-censoring OPJM exhibits a similar bias as in scenario 1.

3.2.3 | Scenario 3 – True model: Left-censoring OPJM

The convergence rate of the C-TPJM (73%) is low, this is due to the data generating mechanism that gives unstable parameter estimates in the binary part (the probability of positive value at baseline is close to 1, corresponding to a linear predictor that converges towards $+\infty$ with the logit link function). Fixing the intercept to a reasonable value of 6.0 solves this issue while not changing the parameters estimates. As expected, the M-TPJM gives unbiased values for the continuous part (Table 3), although we notice slightly lower coverage probabilities for the fixed slope effect ($\hat{\beta}_1 = -0.55$, SD = 0.06, CP = 83%), the interaction of the slope and treatment ($\hat{\beta}_3 = 0.35$, SD = 0.08, CP = 86%) and the

error term ($\hat{\sigma}_\varepsilon = 0.30$, $SD = 0.01$, $CP = 75\%$). In the survival part, all the models are again unbiased, with similar precision and coverage. The random intercept and slope are properly estimated but their correlation is slightly biased upwards with the M-TPJM ($\hat{corr}_{b_0b_1} = 0.44$, $SD = 0.18$, true value is $corr_{b_0b_1} = 0.2$) while the C-TPJM finds no correlation between the intercept and slope ($\hat{corr}_{b_0b_1} = -0.01$, $SD = 0.21$). We also notice a strong correlation between the random intercept from the binary and continuous parts for both the C-TPJM ($\hat{corr}_{ab_0} = 0.93$, $SD = 0.04$) and the M-TPJM ($\hat{corr}_{ab_0} = 0.94$, $SD = 0.04$), this is due to the data generating mechanism where censored values are the 10% smallest observed values, thus inducing a strong correlation between the mean value of the biomarker and the probability of positive value.

3.2.4 | Scenario 4, 5, and 6 – increased zero rate

We investigated three similar simulation scenarios with an increased rate of zeros (20%). The conclusions are similar to our first set of simulations (scenarios 1–3), exhibiting strong bias for the left-censoring OPJM and moderate bias for the C-TPJM and M-TPJM when they are mis-specified (see Tables S1–S3). However, these biases increased compared to the three firsts simulations scenarios, indicating that the excess of zeros impacts substantially parameter estimates.

TABLE 3 Summary of the results of simulations scenario 3 (true model: Left-censoring OPJM), 300 datasets with 400 individuals each and 1000 integration points, 10.04% zeros on average ($SD = 0.02$). The true value of the parameters estimated in the continuous part of the C-TPJM are unknown, therefore coverage probabilities are not provided for these parameters.

Variable	Left-censoring OPJM		C-TPJM	C-TPJM	M-TPJM
	Est. ^a	(SD) ^b [CP] ^c	Est. (SD) [CP]	Est. (SD) [CP]	Est. (SD) [CP]
Binary part					
Intercept	α_0		7.89 (0.78)	6.00 (fixed)	5.95 (0.81)
Time	α_1		−3.50 (0.55)	−2.73 (0.34)	−2.47 (0.47)
Treatment	α_2		2.52 (0.90)	2.96 (0.83)	1.60 (0.62)
Time: treatment	α_3		1.18 (0.76)	0.63 (0.63)	0.72 (0.49)
Continuous part					
Intercept	$\beta_0 = 1.5$	1.52 (0.05) [93%]	1.54 (0.04)	1.50 (0.05)	1.51 (0.05) [96%]
Time	$\beta_1 = -0.5$	−0.51 (0.05) [94%]	−0.42 (0.05)	−0.44 (0.05)	−0.55 (0.06) [83%]
Treatment	$\beta_2 = 0.3$	0.30 (0.06) [94%]	0.29 (0.06)	0.34 (0.06)	0.32 (0.07) [93%]
Time: treatment	$\beta_3 = 0.3$	0.31 (0.08) [92%]	0.23 (0.07)	0.24 (0.07)	0.35 (0.08) [86%]
Residual S.E.	$\sigma_\varepsilon = 0.3$	0.30 (0.01) [93%]	0.29 (0.01)	0.29 (0.01)	0.30 (0.01) [75%]
Survival part					
Treatment	$\gamma = -0.2$	−0.21 (0.13) [95%]	−0.21 (0.13) [95%]	−0.20 (0.13) [94%]	−0.21 (0.13) [95%]
Association	$\varphi = 0.08$	0.08 (0.02) [92%]	0.08 (0.02) [93%]	0.08 (0.02) [93%]	0.08 (0.02) [93%]
Random effects					
Intercept (binary part)	σ_a		4.53 (0.44)	3.62 (0.27)	2.82 (0.42)
Intercept (continuous part)	$\sigma_{b_0} = 0.6$	0.60 (0.03)	0.59 (0.03)	0.59 (0.03)	0.62 (0.03)
Slope (continuous part)	$\sigma_{b_1} = 0.3$	0.30 (0.05)	0.21 (0.05)	0.22 (0.05)	0.29 (0.06)
	$corr_{ab_0}$		0.93 (0.04)	0.93 (0.04)	0.94 (0.04)
	$corr_{ab_1}$		0.33 (0.21)	0.33 (0.21)	0.69 (0.13)
	$corr_{b_0b_1} = 0.2$	0.22 (0.18)	−0.01 (0.21)	−0.02 (0.21)	0.44 (0.18)
Convergence rate		100%	73%	100%	100%

^aMean of parameter estimates.

^bStandard deviation from the mean.

^cCoverage probability.

3.2.5 | Conclusion

To conclude, the left-censoring OPJM gives biased estimates of the mean biomarker value and evolution over time when the true model is either the C-TPJM or the M-TPJM. The M-TPJM provides an accurate estimate of the biomarker trajectory under scenarios 1 and 3 but not under scenario 2, as expected (see Figure 2). In our scenarios, the association between the biomarker and the survival was driven largely by early follow-up (where censorship rate is low), thus the parameter quantifying this association was not affected by the bias of the mean biomarker value observed in the late follow-up (see Figure 2). The assumption of independence between the binary and continuous parts conditionally on the random effects with the C-TPJM can result in unstable parameter estimations and convergence issues when this assumption does not hold, as observed in scenario 3. Finally, the C-TPJM is not able to recover the correlation between the random effects when it is not the true model while the M-TPJM gives a good approximation of this correlation structure in all the scenarios.

4 | APPLICATION TO METASTATIC HEAD AND NECK CANCER DATA

4.1 | Data

The SPECTRUM study (Study of Panitumumab Efficacy in Patients With Recurrent and/or Metastatic Head and Neck Cancer) consists of a phase 3 randomized clinical trial (RCT) of chemotherapy with or without panitumumab in patients with metastatic and/or recurrent squamous cell carcinoma of the head and neck (SCCHN). The objective of the study is to compare the treatment effect of panitumumab in combination with chemotherapy versus chemotherapy alone as first line therapy for metastatic and/or recurrent SCCHN. This dataset is freely available on [ProjectDataSphere.org](https://projectdatasphere.org) (Project Data Sphere is an initiative to provide access to individual patient data from RCTs across numerous cancer types from industry and academia).

Between May 15, 2007 and March 10, 2009, 657 patients were randomly assigned (327 to the panitumumab group and 330 to the control group). The inclusion of the patients started at the date of randomization. The data for analysis includes a subset of 449 patients (i.e., 137 patients excluded from the publicly available dataset and out of them, 71 had no biomarker measurements). The median overall survival (OS) is 0.61 years for the control group (arm A) and 0.81 for the panitumumab group (arm B), 370 patients (82%) died during follow-up. There are 1913 repeated measurements of the SLD, 161 of which are zero values (8%). The number of individual repeated measurements for this biomarker varies between 1 and 29 with a median of 4. The main conclusion of the trial was that the addition of panitumumab to chemotherapy did not improve the OS but it improved the progression-free survival (PFS) and had an acceptable toxicity.¹⁵ However a better PFS does not always lead to improved OS.¹⁶

We chose 5 knots for the splines approximating the baseline hazard function based on an univariate survival model. The penalization term, found with cross-validation, is $\kappa = 0.02$. We use 2000 Monte-Carlo integration points for the numerical approximation of the integral over the random effects.

4.2 | Results from the M-TPJM

In this RCT, there is no zero value at baseline as all patients have at least one measurable lesion at inclusion. For that reason, the estimation of the intercept in the binary part of the conditional two-part model is unstable and led to convergence issues. We therefore decided to fix the intercept value at 8.0 for both the conditional and marginalized models, which corresponds to a baseline mean probability of zero value of 3×10^{-4} . This value was chosen by running the M-TPJM without fixing this intercept while the C-TPJM had convergence issue, other values (e.g., 6.0 or 10.0) also led to convergence issues with the C-TPJM. The results of the M-TPJM with the CL and the SRE association are presented in Table 4. We accounted for a treatment effect at baseline and an interaction between treatment and time in both the binary and continuous parts of the two-part model. The baseline effect was included to account for a possible bias in randomization since the publicly available dataset used in our analysis is just a subset of the full original randomized clinical trial.

4.2.1 | Binary part

The treatment difference in the linear predictor at baseline is negative and slightly significantly different from zero (i.e., p -value barely under 0.05) with the CL association ($\hat{\alpha}_{trt} = -1.00$, SE = 0.44) and the SRE association ($\hat{\alpha}_{trt} = -0.94$, SE = 0.51). In a RCT, a significant difference between treatments arms at baseline could result from a bias in randomization (i.e., since we only used a subset of individuals), a lack of flexibility in the function describing the evolution of the outcome over time or simply by chance, as we'd expect to see this 5% of the time. The slope effect of time is negative and highly significant with the CL ($\hat{\alpha}_{time} = -3.67$, SE = 0.37) and the SRE associations ($\hat{\alpha}_{time} = -3.38$, SE = 0.47). This means that the probability of observing a zero value (i.e., complete remission of the measured tumors) increases over time for the reference treatment (arm A). The time by treatment interaction effect on the probability of observing a positive value is significantly negative for both the CL ($\hat{\alpha}_{time \cdot trt} = -2.02$, SE = 0.49) and SRE associations ($\hat{\alpha}_{time \cdot trt} = -1.54$, SE = 0.57), meaning that the patients receiving treatment arm B are associated with higher odds of zero value over time compared to patients receiving treatment arm A.

4.2.2 | Continuous part

The marginal mean value of the SLD at baseline is found similar between the two treatment arms ($\beta_0 \simeq 1.4$). The slope effect of time is negative and significant (CL: $\hat{\beta}_{time} = -0.68$, SE = 0.08 and SRE: $\hat{\beta}_{time} = -0.61$, SE = 0.07). This effect can be interpreted as a multiplicative time effect on the marginal mean of the biomarker given by $\exp(-0.68) = 0.51$ for the CL association (respectively $\exp(-0.61) = 0.54$ for the SRE association). This corresponds to a reduction of 49% of the SLD value per year among patients receiving the reference treatment (arm A) with the CL association model (respectively a reduction of 46% per year with the SRE association model). The two M-TPJMs do not find a significant treatment effect at baseline nor time by treatment interaction, therefore patients receiving treatment arm B have a similar decreasing trend of SLD over time than those in arm A.

4.2.3 | Survival part

The interpretation of the covariate effects in the survival part depends on the association structure specified for the TPJM. With the SRE association, the parameters are interpreted in terms of effect on the risk of death accounting for some individual heterogeneity of the population (as specified by the random effects). The effect of treatment on the risk of death, independent of the biomarker, is not significantly different from zero neither with the SRE association structure ($\hat{\gamma} = -0.07$, SE = 0.11) nor with the CL association ($\hat{\gamma} = -0.05$, SE = 0.11). Regarding a possible effect of treatment through its association with the biomarker, the treatment and treatment by time interaction are not significantly associated with the mean of the biomarker under either the CL or SRE association even though the probability of a positive SLD value is decreasing with time and at a higher rate for treatment arm B vs. A. For the SRE association, the association between the individual heterogeneity at baseline (random intercept from the continuous part) and the risk of death is positive and slightly significant ($\hat{\phi}_{b_0} = 0.42$, SE = 0.19), indicating that the baseline value of the SLD is predictive of the risk of death. However, for the CL association, the current value of the biomarker is positively and very significantly associated with the terminal event ($\hat{\phi} = 0.08$, SE = 0.01), indicating that the risk of death increases with the value of the SLD. As illustrated in Figure S1, the model shows no difference in the survival curves according to treatment arm. The other advantage of the M-TPJM with the CL association, is that it allows to quantify the effect of one unit increase in the biomarker on the risk of terminal event. For instance, the hazard ratio of a patient with a 1 cm increase in the SLD value is associated with an increased risk of death of 8% ($\exp(0.08) = 1.08$). The M-TPJM with the SRE association can be helpful to characterize how individuals who deviate by a certain amount from the mean SLD trajectory (e.g., 1 standard deviation of the baseline biomarker value) have an increased risk of terminal event compared to a patient with an average SLD profile. For instance, let us assume a clinician is interested in the top 15% patients who had the largest SLD value at baseline compared to the average patient. Their random effect b_{0i} should be higher than 1 standard deviation, that is from Table 4, $b_{0i} > 0.42$. Conditional on $b_{0i} > 0.42$, the mean values of the random effects can be derived by sampling from a conditional multivariate normal distribution with correlation matrix given in Table 4. These conditional means are 3.66, 0.90 and 0.48 for a , b_0 and b_1 , respectively. Therefore, these top 15% individuals increase their chance to have the terminal event (i.e., to die) measured by an hazard ratio of

TABLE 4 Application to metastatic and/or recurrent squamous-cell carcinoma of the head and neck data.

Model	Left-censoring OPJM		Conditional TPJM		Marginalized TPJM	
	Current level	Shared random effects	Current level	Shared random effects	Current level	Shared random effects
Association	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)
Binary part (SLD > 0 vs. SLD = 0)						
Intercept	α_0	8.00 (fixed)	8.00 (fixed)	8.00 (fixed)	8.00 (fixed)	8.00 (fixed)
Time (year)	α_1	-3.41*** (0.78)	-3.41*** (0.78)	-3.27* (0.95)	-3.67*** (0.37)	-3.38*** (0.47)
Treatment (B/A)	α_2	-0.95 (0.74)	-0.95 (0.74)	-0.40 (0.89)	-1.00* (0.44)	-0.94 [‡] (0.51)
Age (> 65 years)	α_3	-1.14 (0.81)	-1.14 (0.81)	-1.77* (0.79)	1.58*** (0.37)	-0.69 (0.46)
Sex (M/F)	α_4	2.74*** (0.83)	2.74*** (0.83)	2.58** (0.84)	2.46*** (0.47)	2.06** (0.68)
Time: treatment (B/A)	α_5	-1.84** (0.71)	-1.84** (0.71)	-1.83 (1.31)	-2.02*** (0.49)	-1.54** (0.57)
Continuous part						
Intercept	β_0	1.46*** (0.10)	1.46*** (0.10)	1.47*** (0.09)	1.37*** (0.08)	1.37*** (0.09)
Time (years)	β_1	-0.87*** (0.10)	-0.86*** (0.10)	-0.62*** (0.07)	-0.68*** (0.08)	-0.61*** (0.07)
Treatment (B/A)	β_2	-0.05 (0.06)	-0.04 (0.06)	-0.04 (0.06)	-0.03 (0.06)	-0.01 (0.06)
Age (>65 years)	β_3	0.11 (0.08)	0.11 (0.08)	0.09 (0.08)	0.10 (0.07)	0.03 (0.08)
Sex (M/F)	β_4	0.32*** (0.09)	0.32*** (0.09)	0.30*** (0.10)	0.38*** (0.08)	0.38*** (0.09)
Time: treatment (B/A)	β_5	-0.06 (0.14)	-0.05 (0.14)	-0.10 (0.09)	-0.12 (0.12)	-0.13 (0.11)
Residual S.E.	σ_ϵ	0.41 (0.01)	0.41 (0.01)	0.31 (0.01)	0.30 (0.01)	0.30 (0.01)
Death risk						
Treatment (B/A)	γ_1	-0.06 (0.11)	-0.04 (0.11)	-0.05 (0.11)	-0.05 (0.11)	-0.07 (0.11)
Age (>65 years)	γ_2	0.18 (0.14)	0.29* (0.14)	0.20 (0.14)	0.18 (0.14)	0.23 (0.14)
Sex (M/F)	γ_3	0.14 (0.17)	0.28 (0.17)	0.14 (0.17)	0.14 (0.17)	0.29 (0.17)
Association						
$E[Y_{ij}]$	φ	0.08*** (0.01)	0.08*** (0.01)	0.08*** (0.01)	0.08*** (0.01)	0.08*** (0.01)
Binary part intercept	φ_a			0.04 (0.03)		0.00 (0.07)
Continuous part intercept	φ_{b_0}		0.48*** (0.11)	0.39*** (0.12)		0.42* (0.19)
Continuous part slope	φ_{b_1}		0.16* (0.06)	0.06 (0.16)		0.30 (0.23)
Random effects						
Intercept (binary part)	σ_a			4.65		4.47
Intercept (continuous part)	σ_{b_0}	0.59	0.59	0.62	0.61	0.63
Slope (continuous part)	σ_{b_1}	1.51	1.44	0.72	0.99	1.01

(Continues)

TABLE 4 (Continued)

Model	Left-censoring OPJM		Conditional TPJM		Marginalized TPJM	
	Current level	Shared random effects	Current level	Shared random effects	Current level	Shared random effects
Association	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)
$corr_{ab}$		0.12	0.37	0.56	0.59	
$corr_{ab}$			0.52	0.36	0.66	0.80
$corr_{b_0b_1}$	0.26	0.26	-0.06	0.13	0.25	0.31
LCV	1.9790	1.9803	1.0525	1.0524	1.0072	1.0082
Computation time (seconds)	1417	160	5271	2019	3870	908
With MPI parallel computing over 80 CPUs (Intel Xeon E5-4627 v4 2.60 GHz)						

^a $p < 0.1$;

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

HR = $\exp(0.00 \cdot 3.66 + 0.42 \cdot 0.90 + 0.30 \cdot 0.48) = \exp(0.52) = 1.69$, 95%CI = 1.39 – 2.02, compared to a patient who has an average longitudinal SLD profile. The confidence intervals were obtained by sampling parameters from the Hessian matrix.

4.3 | Comparison of the M-TPJM with the left-censoring OPJM

The difference in the treatment of the zero values leads to a steeper decrease of the mean biomarker value over time for the left-censoring OPJM (CL: $\hat{\beta}_{time} = -0.87$, SE = 0.10, SRE: $\hat{\beta}_{time} = -0.86$, SE = 0.10) compared to the M-TPJM (CL: $\hat{\beta}_{time} = -0.68$, SE = 0.08, SRE: $\hat{\beta}_{time} = -0.61$, SE = 0.07). The residual error is higher with the left-censoring OPJM ($\hat{\sigma}_\epsilon = 0.41$, SE = 0.01) than with the M-TPJM ($\hat{\sigma}_\epsilon = 0.30$, SE = 0.01), indicating a better fit of the latter. Our results are therefore in line with our simulations when OPJM is not the true model. Overall, the standard errors of the parameter estimates in the biomarker model are lower under the M-TPJM than the OPJM. Nonetheless, the SRE association shows a similar but more significant relationship between the random intercept and slope of the mean biomarker value and the risk of death with the left-censoring OPJM ($\hat{\varphi}_{b_0} = 0.48$, SE = 0.11, $\hat{\varphi}_{b_1} = 0.16$, SE = 0.06) compared to the M-TPJM ($\hat{\varphi}_a = 0.00$, SE = 0.07, $\hat{\varphi}_{b_0} = 0.42$, SE = 0.19, $\hat{\varphi}_{b_1} = 0.30$, SE = 0.23), likely due to the better fit of the latter model. In particular, the random slope standard deviation is found higher with the left-censoring OPJM (CL: $\hat{\sigma}_{b_1} = 1.51$, SRE = $\hat{\sigma}_{b_1} = 1.44$) compared to the M-TPJM (CL: $\hat{\sigma}_{b_1} = 0.99$, SRE: $\hat{\sigma}_{b_1} = 1.01$).

4.4 | Comparison of the M-TPJM with the C-TPJM

The LCV criterion indicates that the M-TPJM fits the data better than the C-TPJM for each association structure. As proposed in Commenges et al.,¹³ the comparison of the LCV value can be classified according to the order of the difference. A difference of order 10^{-1} , 10^{-2} , 10^{-3} , and 10^{-4} may be qualified as ‘large’, ‘moderate’, ‘small’ and ‘negligible’, respectively. The difference in the LCV value between the M-TPJM (CL: LCV = 1.0072, SRE: LCV = 1.0082) and the C-TPJM (CL: LCV = 1.0525, SRE: LCV = 1.0524) is moderate in favor of the M-TPJM and the difference between the CL and the SRE association structures is small in favor of the CL with the left-censoring OPJM (CL: LCV = 1.9790, SRE: LCV = 1.9803) and the M-TPJM where it is negligible with the C-TPJM. We plotted the mean biomarker trajectory estimated by the left-censoring OPJM, the C-TPJM and the M-TPJM in Figure S3. As observed in the simulations when the M-TPJM is the true model, the C-TPJM tends to over-estimate the probability of zeros over time.

In line with the simulation results when M-TPJM is the true model, the variability of the parameter estimates in the binary part is lower under the M-TPJM than the C-TPJM. Treatment arm B (chemotherapy + panitumumab) vs. arm A (chemotherapy alone) is associated with a more significant reduction in the probability of positive value over time with the M-TPJM (CL: $\hat{\alpha}_{time \cdot trt} = -2.02$, SE = 0.49, SRE: $\hat{\alpha}_{time \cdot trt} = -1.54$, SE = 0.57) compared to the C-TPJM with the CL association structure ($\hat{\alpha}_{time \cdot trt} = -1.84$, SE = 0.71) and the SRE association ($\hat{\alpha}_{time \cdot trt} = -1.83$, SE = 1.31). In the continuous part, the effect of treatment is not found significantly different from zero under either the C-TPJM or the M-TPJM but its interpretation is different under these 2 models, as illustrated in Figure 1. The M-TPJM finds no treatment effect on the overall mean biomarker value (CL: $\hat{\beta}_{time \cdot trt} = -0.12$, SE = 0.12, SRE: $\hat{\beta}_{time \cdot trt} = -0.13$, SE = 0.11) where the C-TPJM finds no treatment effect on the biomarker value conditional on a positive value (CL: $\hat{\beta}_{time \cdot trt} = -0.10$, SE = 0.09, SRE: $\hat{\beta}_{time \cdot trt} = 0.07$, SE = 0.09). As found in our simulation results, the effect of treatment independent of the biomarker and the association parameters are very similar between the M-TPJM and C-TPJM.

To conclude, as observed in our simulation results, the C-TPJM can lead to biased estimates and incorrect statistical inference when it is not the true model (the M-TPJM is the best model in the application).

An advantage of the M-TPJM is that it is straightforward to evaluate the hazard ratio of treatment at a given time point with the current value association structure compared to the C-TPJM (note that with the shared random effects, this hazard ratio is directly given by exponentiating the treatment effect in the survival submodel for both the M-TPJM and the C-TPJM). For example, from Table 4, we can calculate that the hazard ratio of treatment for the reference individual (i.e., for a woman 65 years old or less), 1 year after randomization, is 0.93, 95%CI = 0.75 – 1.16. The decrease in the SLD due to treatment effect is reflected in the hazard, leading to a slightly improved survival over time, although not statistically significant. This hazard ratio can also be obtained from the C-TPJM but this requires combining the effect of treatment on the binary and on the continuous parts of the biomarker model to obtain its effect on the marginal mean of the biomarker while it is directly given by the M-TPJM. For instance, the hazard ratio of treatment, 1 year

after randomization with the C-TPJM is $0.93, 95\%CI = 0.75 - 1.17$. The details of these computations are given in the supplementary material. The computation time of the M-TPJM (CL: 3870 s, SRE: 908 s) is reduced compared to the C-TPJM (CL: 5271 s, SRE: 2019 s) while it is higher than the left-censoring OPJM (CL: 1417 s, SRE: 160 s) because of the additional parameters included for the binary part.

5 | DISCUSSION

We proposed a marginalized two-part joint model for a longitudinal semicontinuous biomarker and a terminal event that allows to obtain directly the effect of covariates, such as treatment effect, on the marginal mean of the biomarker. This M-TPJM is as an alternative to the conditional two-part joint model. While the mean biomarker value at baseline and over time is directly estimated from the M-TPJM, it is obtained from the mixing distributions of the zero and non-zero components in the C-TPJM, which imposes a non-linear curve for the mean biomarker value over time, not always justified. The effect of covariates on the marginal mean of the biomarker under the C-TPJM is also not directly available. We also proposed for the M-TPJM two association structures to link the biomarker to the risk of terminal event. The first one, the current value association, allows to explore time-dependent effect of covariates on survival through the biomarker, as well as the effect of covariates on the terminal event independent of the biomarker. The second one consists of sharing only the random effects from the two-part model, which evaluates the relationship between the risk of terminal event and the individual deviation from the population mean of the biomarker, including for example the baseline odds of a positive value, baseline value and slope for the whole trajectory. An illustration of these 2 association structures in the SPECTRUM study is given at the end of 4.2 and 4.4. In the presence of true zeros, if the clinical interest is in the association between the marginal mean of the biomarker and a terminal event, the M-TPJM should be favored over the C-TPJM. In contrast, if the interest lies in the positive values of the biomarker (i.e., zeros excluded) and the probability of positive values and their association with a terminal event, the C-TPJM must be used. Indeed, the C-TPJM allows to link the binary and continuous parts of the biomarker model to the risk of event separately (see Rustand et al.²), making it possible to have information and make decisions for the subpopulation of patients with a zero value or for those with a positive value, separately. Therefore these two formulations of the two-part joint model are complementary and can answer different clinical questions of interest.

Our simulation studies shows marked differences across the three models applied: the left-censoring OPJM, the C-TPJM and the M-TPJM. The left-censoring OPJM was severely biased in the estimation of treatment effect on the biomarker when true zero values (i.e., not censored) were present. The C-TPJM can account for excess of zeros but led to biased estimates and wrong inference about treatment effect on the marginal mean value of the biomarker whenever it was not the true model. In addition, the C-TPJM could have convergence issues due to the assumed independence between the probability of zero and the expected value among positive conditional on the random effects. The M-TPJM provided an accurate inference about the biomarker and covariate effects on the biomarker in most situations, unless the distribution of the biomarker over time is not linear on the log scale. In this case, smooth functions of time should be included in the longitudinal part.

The differences observed across models did impact the inference about the effect of treatment captured through the biomarker on the terminal event and shared through the current value association but to a lesser extent, the association of the treatment on the terminal event independent of the biomarker and the association between the biomarker and the terminal event. This could be the consequence of the heavy censoring present in the simulated data (which mimicked the real data) and the fact that the estimated mean value of the biomarker during the early follow-up was relatively close across models (see Figure 2). In other situations with lower censoring rate or higher proportion of zeros, it is not excluded that the treatment effect on the risk of terminal event not captured by the biomarker and association parameter(s) be also affected by the model assumptions.

Our application to a cancer clinical trial assessing two treatment arms for squamous cell carcinoma of the head and neck illustrates the interest of the M-TPJM. Indeed, clinicians are often interested in population average estimates of covariates, such as treatment, to make decision. We recall that the original trial concluded that the addition of panitumumab to chemotherapy did not improve OS but led to better progression-free survival (Vermorken et al.¹⁵). Based on LCV criterion, the M-TPJM fitted the data better than the C-TPJM. In line with our simulation results when the M-TPJM was the true model, the C-TPJM is associated with higher uncertainty in the binary part compared to the M-TPJM, which in turn yields higher uncertainty on the mean biomarker trajectory. Specifically, the C-TPJM found

an attenuated treatment effect compared to the M-TPJM for the disappearance of all target lesions over time. Finally, the M-TPJM did not conclude to a treatment effect on the overall mean of the biomarker either at baseline or during the follow-up.

This work has several limitations. For instance, in the cancer clinical trial application, the SLD measures the longest diameter of target lesions, which can be subject to important measurement error. It could be more appropriate to use instead a more accurate measurement such as the total volume of the tumors, although it is usually unavailable in clinical trials as it is not part of the RECIST criteria. In this work, the M-TPJM was not developed specifically to capture a non-linear mean biomarker trajectory on the log scale. The inclusion of time-dependent covariates and interactions as well as non-linear functions of time could account for such trajectories. Besides, an extension of the marginalized two-part model has been proposed using a generalized Gamma distribution (that includes the log-normal as a specific case) to link the outcome to the linear predictor in the continuous part and allows more flexibility in the biomarker trajectory but has not yet been developed for joint models.^{17,18} The short follow-up of patients and low proportion of zero values did not allow more flexible functions of time to be included in either the binary or continuous part of the model, as it could lead to unstable parameters estimation and convergence issues. However we were able to illustrate the differences between the C-TPJM and the M-TPJM in terms of the mean biomarker trajectory and its association with the risk of death. In this work, we only evaluated multiplicative effect of covariates on the conditional and marginal mean biomarker value, although we could have evaluated the additive effect of covariates on the log scale by fitting a linear mixed effects model on log-transformed biomarker values as proposed in Rustand et al.² This model formulation can also be interpreted in terms of change in the median of the response variable by exponentiating the parameter estimates but it does not offer uncertainty measure nor statistical significance on the original scale of the response variable. Alternative algorithms such as EM or BFGS could be used to fit the TPJMs although they have not been found to improve the parameter estimation compared to the Levenberg–Marquardt's algorithm.¹⁹ Moreover, a perspective to improve our approach could be to use importance sampling to draw Monte Carlo samples when integrating out the density of random effects in the likelihood instead of using the standard random samples. Finally, the random-effects joint models described in this article rely on the assumption that the biomarker and the time-to-event are independent given the random effects but when this assumption does not hold, the joint model can provide biased estimates. This will be the case under MAR mechanisms of dropout.²⁰ This could occur in our application if the time of dropout was a MAR missing dropout mechanism, that is when death is only predictable by observed tumor volumes. In this context mixed models are more robust than joint models.

The use of joint modeling is a powerful tool to answer questions of interest in medical studies when a longitudinal marker and survival times are of interest as well as their relationship.⁷ The semicontinuous nature of the biomarker is usually handled by either left-censoring that assumes an often unrealistic assumption of no true zeros observed or a conditional two-part model that complicates the marginal interpretation of covariates effect. In this article, we introduced the M-TPJM and contrasted it with alternative modeling strategies (i.e., OPJM, C-TPJM, each with different association structures) and described the strengths and limitations of each approach in terms of clinical decision-making. Indeed, this family of methods offers various ways to assess the effectiveness of drugs, which might be all of clinical interest. For instance, in the context of clinical trials, the M-TPJM can be used to compare the longitudinal semicontinuous biomarker values and terminal event risk between the treatment and control groups as illustrated in our application section. It allows a direct estimation of the effect of covariates, including treatment effect, on the marginal mean of the biomarker, which is of major interest for clinicians and decision-makers. When the M-TPJM indicates a significant improvement in the biomarker value or a reduction in the risk of the terminal event for the treatment group, this suggests that the drug is effective in treating the disease in the whole group of patients studied. Treatment interactions with covariates can be included in the model to evaluate the effectiveness of treatment for subgroups of patients described by those covariates. The use of patients' data with appropriate models that might predict response to a particular medication is important as this information can help clinicians and healthcare providers make more personalized treatment decisions for their patients, improving overall health outcomes. Moreover, we argue that the C-TPJM and the M-TPJM complement each other as the former is able to assess subgroups that may respond differently to the treatment by separating individuals with zero values for the biomarker from those with positive values in the hazard function while the M-TPJM offers a population average overview. The M-TPJM is an useful tool to account for the excess of zeros while keeping the interpretation on the entire population instead of subpopulations, therefore enhancing decision-making for clinicians. Beyond the application to solid tumor cancer data, we propose a freely available software (*frailtypack*, Król et al.²¹) that can be applied to several other situations that include a longitudinal semicontinuous

biomarker and survival times (covariates measuring symptoms of a disease or quantifying exposure are often semicontinuous).

ACKNOWLEDGMENTS

This publication is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this publication.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The SPECTRUM Study dataset is freely available on ProjectDataSphere.org (Project Data Sphere is an initiative to provide access to individual patient data from RCTs across numerous cancer types from industry and academia).

ORCID

Denis Rustand  <https://orcid.org/0000-0001-9708-5220>

REFERENCES

- Liu L. Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Stat Med*. 2009;28(6):972-986.
- Rustand D, Briollais L, Tournigand C, Rondeau V. Two-part joint model for a longitudinal semicontinuous marker and a terminal event with application to metastatic colorectal cancer data. *Biostatistics*. 2022;23(1):50-68.
- Olsen MK, Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. *J Am Stat Assoc*. 2001;96(454):730-745.
- Smith VA, Neelon B, Preisser JS, Maciejewski ML. A marginalized two-part model for longitudinal semicontinuous data. *Stat Methods Med Res*. 2017;26(4):1949-1968.
- Smith VA, Preisser JS, Neelon B, Maciejewski ML. A marginalized two-part model for semicontinuous data. *Stat Med*. 2014;33(28):4891-4903.
- Shahrokhbadi MS, Chen D-G, Mirkamali SJ, Kazemnejad A, Zayeri F. Marginalized two-part joint modeling of longitudinal semi-continuous responses and survival data: with application to medical costs. *Mathematics*. 2021;9(20):2603.
- van Eijk RP, Roes KC, van den Berg LH, Lu Y. Joint modelling of endpoints can be used to answer various research questions in randomized clinical trials. *J Clin Epidemiol*. 2022;147:32-39.
- Tobin J. Estimation of relationships for limited dependent variables. *Econometrica*. 1958;26(1):24-36.
- Smith VA, Maciejewski ML, Olsen MK. Modeling semicontinuous longitudinal expenditures: A practical guide. *Health Serv Res*. 2018;53(S1):3125-3147.
- Smith VA. Marginalized two-part models for semicontinuous data with application to medical costs. PhD thesis, The University of North Carolina at Chapel Hill. 2015.
- van Oudenhoven FM, Swinkels SH, Ibrahim JG, Rizopoulos D. A marginal estimate for the overall treatment effect on a survival outcome within the joint modeling framework. *Stat Med*. 2020;39(13):4120-4132.
- Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. *J Soc Ind Appl Math*. 1963;11(2):431-441.
- Commenges D, Joly P, Gégout-Petit A, Liquet B. Choice between semi-parametric estimators of Markov and non-Markov multi-state models from coarsened observations. *Scand J Stat*. 2007;34:33-52.
- Sylvestre M-P, Abrahamowicz M. Comparison of algorithms to generate event time conditional on time-dependent covariates. *Stat Med*. 2008;27:2618-2634.
- Vermorken JB, Stöhlmacher-Williams J, Davidenko I, et al. Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (spectrum): an open-label phase 3 randomised trial. *Lancet Oncol*. 2013;14(8):697-710.
- Prasad V, Kim C, Burotto M, Vandross A. The strength of association between surrogate end points and survival in oncology: a systematic review of trial-level meta-analyses. *JAMA Intern Med*. 2015;175(8):1389-1398.
- Smith VA, Neelon B, Maciejewski ML, Preisser JS. Two parts are better than one: modeling marginal means of semicontinuous data. *Health Services Outcomes Res Methodol*. 2017;17(3-4):198-218.
- Smith VA, Preisser JS. Direct and flexible marginal inference for semicontinuous data. *Stat Methods Med Res*. 2017;26(6):2962-2965.
- Philipps V, Hejblum BP, Prague M, Commenges D, Proust-Lima C. Robust and efficient optimization using a Marquardt-Levenberg algorithm with R package marqLevAlg. arXiv:2009.03840. 2020.
- Thomadakis C, Meligkotsidou L, Pantazis N, Touloumi G. Longitudinal and time-to-drop-out joint models can lead to seriously biased estimates when the drop-out mechanism is at random. *Biometrics*. 2019;75(1):58-68.

21. Król A, Mauguen A, Mazroui Y, Laurent A, Michiels S, Rondeau V. Tutorial in joint modeling and prediction: a statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *J Stat Softw.* 2017;81(3):1-52.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rustand D, Briollais L, Rondeau V. A marginalized two-part joint model for a longitudinal biomarker and a terminal event with application to advanced head and neck cancers. *Pharmaceutical Statistics.* 2023;1-21. doi:[10.1002/pst.2338](https://doi.org/10.1002/pst.2338)