Joint Longitudinal-Survival Modelling of Childhood Leukemia Using EM Algorithm

Anwar Dakhel Handool Al-Aaidy and Suhad Ali Shaheed Al-Tamimi
Department of Statistics, College of Administration and Economics, Al-Mustansiriyah University, 46167 Baghdad, Iraq
anwardakhelhandool@gmail.com, dr.suhadali@uomustansiriyah.edu.iq

Keywords: Joint Model, Longitudinal and Survival Data, Model Evaluation (AIC, RMSE), EM Algorithm.

Abstract: This Study aims to assess the impact of certain clinical factors on children with leukemia by employing The Joint Model for Longitudinal and survival data, using the Expectation–Maximization (EM) algorithm for parameter estimation. In this research, the longitudinal data are represented by repeated measurements of hemoglobin (Hb) levels over time, while the survival data correspond to the follow-up time until the realization of the event (death or censoring). The results revealed that Hb levels significantly decrease over time ($p < 0.0001$). Female patients had lower Hb levels compared to males ($p=0.006$), while older age was associated with slightly higher Hb values ($p=0.0002$). In the survival sub-model, the effects of age and sex were statistically insignificant ($p > 0.05$). Importantly, the negative and significant association parameter (γ, $p=0.0001$) indicated that higher Hb levels reduce the risk of death, highlighting the predictive value of longitudinal information in explaining survival outcomes. Based on the goodness-of-fit criteria (AIC and RMSE), the EM algorithm demonstrated high efficiency in estimating the parameters of the joint model and achieving strong agreement between predicted and observed values. Therefore, the joint model provides a powerful and effective tool for integrating longitudinal and survival information in the study of childhood leukemia, enhancing estimation accuracy and enabling more reliable conclusions about the factors influencing disease progression.

1 INTRODUCTION

Childhood leukemia is one of the most common and serious types of cancer, posing a major challenge to healthcare systems due to its direct impact on survival rates and patients' quality of life. Studying the factors that influence disease progression and survival among affected children is a crucial step toward improving diagnostic and therapeutic strategies. In this context, advanced statistical models have become essential to capture the complexity of medical data, which often consist of repeated longitudinal measurements of biological markers alongside survival outcomes [1], [2].

The Joint Model for Longitudinal and survival data is considered one of the most powerful tools in this field. It integrates longitudinal information (such as hemoglobin levels) with survival outcomes (time-to-event) within a unified framework, providing more accurate estimates and enabling the study of dynamic relationships between biomarkers and the risk of mortality [3]. Several estimation methods have been developed for this model, with the Expectation–Maximization (EM) algorithm being among the most widely used due to its efficiency in handling missing data and complex parameters [4], [5].

This study aims to apply The Joint Model using the EM algorithm to measure the effect of demographic and clinical factors (such as age and sex) on children with leukemia, with a particular focus on the relationship between changes in hemoglobin over time and survival probability. The research also provides interpretation of the estimation results and highlights conclusions that may contribute to supporting medical decision-making and improving healthcare for this vulnerable group of patients.

2 JOINT MODELS

The concept of Joint Models is based on the modeling of two types of data that form two components within the model. This modeling helps to understand how these data depend on each other, leading to better results, more accuracy, less bias, and more accurate

predictions [1], [4]. Both data in a common model are modeled according to two components: the first records changes over time and the second monitors when events (deaths) occur. Wulfsohn and Tsiatis (1997) were the first to describe the joint model, as these models consist of two sub-models [1], [2].

2.1 Longitudinal Sub-Model

The first sub-model is a Mixed Effect sub-model for modeling longitudinal data, which represents repeated measurements of variables, such as measuring vital signs or monitoring the patient's response to treatment over time. It tracks changes and differences between measurements, such as disease progression or treatment effectiveness. Longitudinal data are often modeled using the Linear Mixed-Effect model, which studies fixed effects representing overall trends and random effects representing individual differences, thereby capturing both population-level trends and individual variations [4], [6]. In addition, the analysis of high-dimensional longitudinal data has become increasingly important, where joint modeling provides a robust framework for addressing complexity and ensuring reliable inference [7].

2.2 Survival Sub-Model

The second sub-model represents the modeling of the time of occurrence of events (e.g., death) based on survival data analysis. The time until the occurrence of the event is observed, with examples including the proportional hazards model and accelerated failure time models. These models estimate the risk or timing of an event. The Cox proportional hazard semi-parametric model is the most common approach, as it evaluates how different factors affect the risk of an event without assuming restrictions on event times [2], [8].

2.3 Combined Joint Model

Joint models integrate both sub-models by allowing the survival component (event occurrence time) to depend on some characteristics of the longitudinal sub-model through various methods [1], [5]. A key feature of joint models is that longitudinal data from a given time until the event is modeled simultaneously with a conditional joint density function, rather than treating the two sub-models as independent [4], [9].

To formulate the joint model, some terms must be defined:

- T_i^*: denotes the true occurrence time of the event for subject *i*;
- T_i: represents the observed time for subject *i*;
- δ_i : is the event index equal to (1) when the real event occurs;
- y_i: is the longitudinal response variable;
- The formulation of the joint model generally involves three stages [1], [10].

2.1 Constructing the Survival Data Sub Model

Let T_i^* be the true event time for subject $(i = 1, 2, \dots, n)$, and let T_i denote the observed survival time. If the subject does not experience the event during the study period, T The observation for subject i is right-censore. Let C_i be the potential censoring time such that:

$$T_i = min(T_i^*, C_i).$$

Define the event indicator as:

$$\delta_i = \begin{cases} 1 \text{ if the true event is observed;} \\ 0 \text{ if the event is right} - \text{censored.} \end{cases}$$

Thus, the observed survival data can be written as: [5], [10]

$$\{(T_i, \delta_i);\ i = 1,2, \dots, n\}.$$

For the i^{th} subject, let $y_i(t)$ denote the observed longitudinal measurement at time point t_{ij}, for *j=1, 2, …,* m_i. Let m_i (t) denote the true longitudinal trajectory of subject i, which is unobserved and differs from $y_i(t)$.

To describe the association between the longitudinal trajectory and the hazard of an event, the proportional hazards model is given by:

$$h_i(t \mid \mu_i(t)) = \lim_{\delta t \to 0} \left\{ \frac{p[t \leqslant T_i^* < t + \delta t \setminus T_i^* \geqslant t, \mu_i(t), w_i]}{\delta t} \right\} = h_0(t)\, exp\{\gamma^\top w_i + \alpha m_i(t)\}, \ t > 0, \quad (1)$$

where:

- $m_i(t)$represents the true values of the time-dependent longitudinal covariates at time *t.*;
- $h_0(t)$ is the unspecified baseline hazard for a subject with all explanatory covariates equal to zero;
- w_i is the vector of baseline explanatory covariates;
- γ is the vector of regression coefficients corresponding to w_i;

- α is the association parameter measuring the effect of the longitudinal data on survival outcomes at time.[4], [10].

The hazard function in (1) assumes that the risk of experiencing the event at time *t*. depends only on the current value of $m_i(t)$, which changes over time.

Based on the relationship between the survival function and the cumulative hazard function, the survival probability can be expressed as:

$$S_i\big(t \setminus M_i(t)\big) = Pr(\tau_i^* > t \setminus \mu_i(t), w_i) = exp\left(-\int_0^t h_0(s)\exp\{\gamma^\top w_i + \alpha M_i(s)\}ds\right). \quad (2)$$

This expression represents the survival function for subject i, conditional on the entire longitudinal process $m_i(t)$ [4], [6]

2.2 Constructing the Longitudinal Sub-Model

The hazard function $h_i\ (t)$ in (1) depends on the true longitudinal outcomes $m_i(t)$ at time t. For each subject, the true longitudinal process is recorded at intermittent time points $\{t_{ij}, j=1, 2,\ldots,m_i\}$, subject to measurement errors. To relate the longitudinal outcomes to the risk of event occurrence, it is necessary to estimate $m_i(t)$.

Therefore, the longitudinal data are modeled using the Linear Mixed-Effects Model (LME). For the i^{th} subject, the observed longitudinal responses are given by:

$$y_{ij}\ (t_{ij}), j=1, 2,\ldots,m_i.$$

Assuming that the longitudinal outcomes follow a normal distribution, the linear mixed-effects model can be specified as follows:

$$y_i(t) = x_i^\tau(t)\beta + z_i^\tau(t)b_i + \epsilon_i(t), \quad (3)$$

$$y_i(t) = m_i(t) + \epsilon_i(t), \quad (4)$$

$$m_i(t) = x_i^\tau(t)\beta + \epsilon_i^\tau(t)b_i. \quad (5)$$

Here, b_i follows a multivariate normal distribution with covariance matrix D, i.e.,

$$b_i \sim N(0, D).$$

The random error term is assumed as $\epsilon_i(t) \sim N(0, \sigma^2)$, and is mutually independent of b_i.

Where:

- $x_i(t)$, β*:* represent the fixed effects;
- $z_i(t), b_i$: represent the random effects;
- $\epsilon_i(t)$: denotes the random error term for subject i at time *t.* [4], [6]

2.3 Constructing the Joint Model

The Joint Model for Longitudinal and survival data consists of two sub-models that are linked together, and all parameters are estimated simultaneously. The Mixed-Effects Model is one of the most commonly used approaches for modeling longitudinal data, while the Proportional Hazards Model is typically employed to handle survival data with time-dependent covariates.

The association between the two sub-models may occur through random effects or regression parameters. To construct The Joint Model for Longitudinal and Survival Data, the following steps are generally taken [5], [11].

1) Specify the trajectory function for the longitudinal process, which correctly represents its evolution over time.
2) Define the structure of the longitudinal component based on the observed values.
3) Specify the appropriate model for the survival component.

The trajectory function plays a central role by representing the true values of the longitudinal covariate across all time points, including unobserved times (free of measurement error). Thus, it allows us to capture both the underlying trajectory of the covariate and its time-dependent effect [9], [12].

The trajectory function can be defined as:

$$W_i(t) = z(t)\ b_i\ \ , t > 0, i = 1,2, \ldots, n \ldots, (6)$$

where:

- $W_i(t)$denotes the trajectory function of subject i at time t, representing the true value of the longitudinal covariate regardless of whether an observation is available at that time;
- $z(t)$ is a vector of continuous functions;
- b_i is the vector of subject-specific random effects (including random intercepts and slopes).

After defining the trajectory, it can be related to the observed longitudinal measurements: [6], [10].

$$y_i = (y_i\ (t_{i1}), y_i\ (t_{i2}), \ldots, y_i\ (t_{iJ_i})$$

with:

$$y_i(t_{ij}) = w_i(t_{ij}) + \epsilon_i(t_{ij})\ \ ;\ i\ = 1, 2, \ldots n. \quad (7)$$
$$j = 1, 2, \ldots J_i = x_i\beta + z_i(t_{ij})^\top b_i + \epsilon_i(t_{ij})$$

where:

- -$y_i(t_{ij})$: observed longitudinal outcome for subject i at time t_{ij};

- $w_i(t_{ij})$: trajectory function, consisting of both fixed and random effects;
- x_i: vector of baseline covariates for subject i;
- β: vector of fixed-effects coefficients;
- z_i: time-dependent design vector;
- b_i: vector of random effects, assumed $b_i \sim N(\mu_b, \Sigma_b)$;
- $\epsilon_i(t_{ij})$: error term, assumed $\epsilon_i(t_{ij}) \sim N(0, \sigma_\epsilon^2)$.

The proportional hazards model can then be extended to incorporate the time-dependent trajectory function:

$$h(u_i) = \lim_{d_u \to 0} \frac{p(u_i \leqslant T_i < u_i + d_u \mid T_i \geqslant u_i, w_i^H(u_i) z_i b_i}{d_u}. \quad (8)$$

$$h(u_i) = h_0(u_i)\, exp\{\eta w_i(u_i) + z_i\theta\} \quad ; \; i = 1,2, \ldots, n.$$

The survival function is then expressed as:

$$\begin{aligned} S(u_i) &= exp\{-H(u_i)\} \\ &= exp\left\{-\int_0^{u_i} h(s)ds\right\} \\ &= ex\,p\{-\textstyle\int_0^{u_i} h_0(s)\exp\{\eta(w_i(s) + z_i\mu\}ds. \end{aligned} \quad (9)$$

Thus, combining the longitudinal and survival components, the joint likelihood function can be written as:

$$\begin{aligned} L(\theta; u_i, \delta_i, y_i, t_i, x_i, z_i) &= \prod_{i=1}^{n} p(u_i, y_i \mid \theta) \\ &= \textstyle\prod_{i=1}^{n} \int [h_0(u_i) ex\, p(\eta\omega_i(u_i) + \\ & z_i\psi)]^{\delta_i} . ex\, p\{-\textstyle\int_0^{u_i} h_0(s) ex\, p(\eta\omega_i(s) + \\ & z_i\psi ds\}. \frac{1}{(2\pi\sigma_\epsilon^2)^{J_i/2}} ex\, p\left\{-\textstyle\sum_{j=1}^{J_i} \frac{\{Y_i(t_{ij}) - W_i(t_{ij})\}^2}{2\sigma_\epsilon^2}\right\}. \\ & p(b_i \mid \mu_b, \Sigma b) db_i. \end{aligned} \quad (10)$$

Under the assumption of full conditional independence, the joint distribution can be factored as:

$$P(y_i, T_i, \delta_i \mid b_i) = p(y_i \mid b_i) \cdot p(T_i, \delta_i \mid b_i). \quad (11)$$

$$P(y_i \mid b_i) = \prod_{j=1}^{m_i} p(y_{ij} \mid b_i). \quad (12)$$

Finally, the survival function conditional on random effects is:

$$S_i(t \mid b_i) = exp\left(-\int_0^t h_0(s) exp\{\gamma^\top w_i + \propto m_i(s)\} ds\right). \quad (13)$$

Hence, the joint likelihood can also be written as:

$$\begin{aligned} P(y_i, T_i, \delta_i) = \int\ & p(y_i \\ & \mid b_i)\{h(T_i \mid b_i)^{\delta_i} S(T_i \\ & \mid b_i)\} p(b_i) db_i. \end{aligned} \quad (14)$$

3 ESTIMATION METHOD FOR JOINT MODEl

After studying the concept of the joint model and identifying all of its components, the estimation methods for the joint model of the combined data will be examined based on equations (13), (14), and (15). In particular, the Maximum Likelihood Estimation (MLE) method will be studied [13].

3.1 Maximum Likelihood Method (MLE)

Wulfsohn and Tsiatis (1997), as well as Rizopoulos (2012), provided a detailed methodology for estimating the joint model using the Maximum Likelihood Estimation (MLE) approach [5].

$$\begin{aligned} & \log L_i(\theta) \\ &= log \int \left\{ \prod_{j=1}^{mi} p(y_{ij} \mid b_{ij}\theta)\{h(T_i \mid b_i; \theta)^{s_i} S_i(T_i \mid b_i; \theta)\} \right. \\ & \qquad \left. p(b_i; \theta) d\, b_i. \right. \end{aligned} \quad (15)$$

Since both integrals do not admit closed-form solutions, numerical approximation methods are required to estimate the joint model. Among these methods, the EM algorithm is widely used [5].

3.1.1 The Expectation–Maximization Algorithm (EM)

The EM algorithm, first developed by Arthur Dempster and colleagues in 1977, is a numerical technique primarily used for parameter estimation in models with incomplete data or latent variables. This algorithm addresses the implications associated with such models that cannot be estimated using traditional statistical methods, making it a pivotal tool for addressing specific challenges in data analysis. Therefore, the EM algorithm is of great importance in the fields of statistics and machine learning.

The purpose of using the EM algorithm is to maximize the likelihood function for models with unobserved data in order to improve prediction [4], [5].

By substantially enhancing statistical parameter estimates, the EM algorithm provides more accurate and reliable results, making it a valuable tool in various research and applied domains.

The basic steps for applying the EM algorithm to the joint model are as follows:

1) Basic equation:
   Represent the Maximum likelihood function for The Joint Model, which the EM algorithm seeks to maximize.

$$\log L_i(\theta) = log \int \left\{ \prod_{j=1}^{mi} p(y_{ij} \mid b_{ij}\theta)\{h(T_i \mid b_i;\theta)^{s_i} S_i(T_i \mid b_i;\theta)\} \right\} p(b_i;\theta) d\, b_i. \quad (16)$$

   Where:
   - y_{ij}: observed longitudinal data;
   - T_i: survival time for each patient;
   - b_i: random effects or latent (unobserved) variables;
   - θ: parameters to be estimated;
   - $h(T_i \mid b_i;\theta)$: conditional hazard function;
   - $S_i(T_i \mid b_i;\theta)$: conditional survival function.

2) E-step (Expectation step):

   Estimate the expected values of the random effects (b_i). Since these random effects are not directly observed, we compute their expectations based on the current parameter estimates:

$$Q(\theta \mid \theta^{(t)}) = E[log\ L_i(\theta) \mid y_{ij}, \tau_i, \theta^{(t)}].$$

   This expectation is obtained from the conditional distribution of the random effects.

3) M-step (Maximization step):
   Update the parameter estimates θ by maximizing the expectation

$$arg\ max_\theta Q(\theta \mid \theta^{(t)}). = \theta^{(t+1)}.$$

4) Iteration until convergence:

   Repeat the E-step and M-step until convergence, i.e., when the difference between successive estimates becomes sufficiently small [4], [5].

$$\theta^{(t+1)} - \theta^{(t)} \approx 0.$$

4 PERFORMANCE MEASURES

To evaluate the predictive performance of the proposed model, several statistical measures are employed. These measures provide a comprehensive assessment of the model's accuracy and efficiency. The most widely used criteria include:

1) The Root Mean Squared Error (RMSE). It is calculated according to the following [8], [10]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y_t})^2}, \quad (17)$$

2) Akaike Information Criterion (AIC). Used to evaluate model quality by balancing goodness of fit and model complexity [4], [10].
   AIC penalizes overfitting by incorporating the number of estimated parameters. It is defined as:

$$AIC = -2\ln(L) + 2k... \quad (18).$$

5 THE PRACTICAL APPLICATION

In this part, The Joint Model for Longitudinal and survival Data was applied using the R software to estimate the parameters and analyze the relationship between longitudinal variables and event time. Estimation methods were employed, and the results were evaluated using performance criteria such as RMSE and AIC to assess the efficiency of the model.

The Figure 1 presents Kaplan–Meier survival curves by sex, showing that the trajectories of males and females are very similar, with no statistically significant difference between the two groups (p=0.8), indicating that sex had no clear effect on survival probability during the follow-up period.

The Figure 2 illustrates the relationship between hemoglobin (Hb) levels and the hazard ratio (HR), showing that higher Hb levels are associated with lower hazard, indicating that increased Hb reduces the likelihood of the event and serves as an important indicator of improved survival.

Table 1 presents the estimation results of the joint longitudinal–survival model using the Expectation–Maximization (EM) algorithm. As shown in Table 1, the longitudinal sub-model indicates that the intercept is positive and highly significant, while time has a negative and statistically significant effect (p = 0.0001), reflecting a gradual decline in hemoglobin (Hb) levels over time. Age shows a positive and significant association with Hb levels (p = 0.0002), whereas female patients exhibit significantly lower Hb values compared to males (p = 0.006).

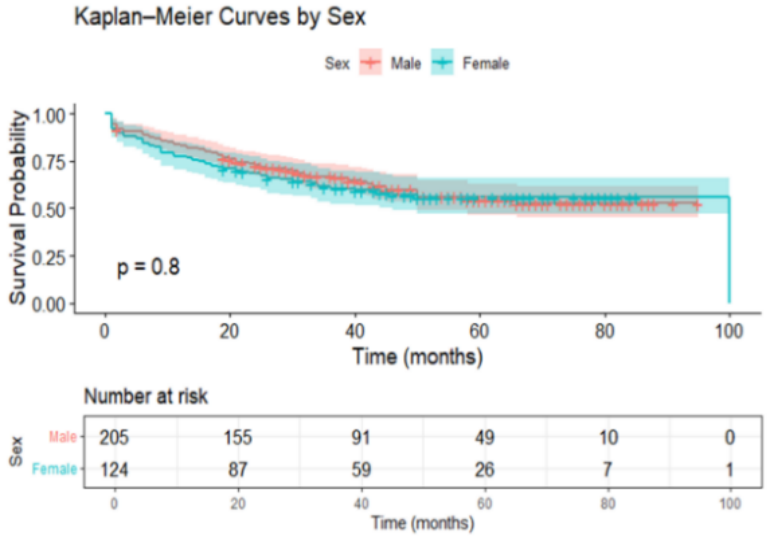

Figure 1: Kaplan–Meier survival curves were analyzed by sex.

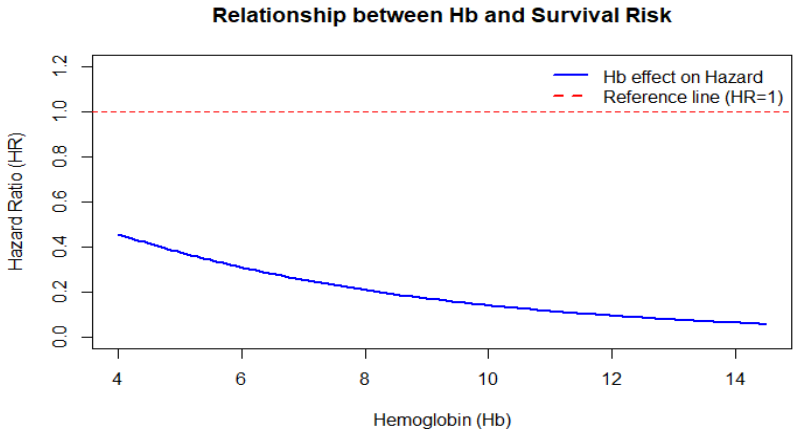

Figure 2: The parameter estimates of the joint model were obtained using the EM algorithm.

Table 1: Estimation Results of the Joint Model (EM).

Model Component	Variable /Parameter	Estimate	Std. Error	Z-value	P-value
Longitudinal sub-model	Intercept	6.54	0.29	22.34	<0.0001
	Time (months)	-0.030	0.0075	-4.04	0.0001
	Age	0.122	0.033	3.68	0.0002
	Female vs Male	-0.666	0.244	-2.73	0.0060
Survival sub-model	Age	-0.045	0.026	-1.75	0.080
	Female vs Male	0.044	0.188	0.24	0.810
Association parameter	γ (gamma)	-0.196	0.049	-4.02	0.0001
Random effects	Intercept variance	4.562	—	—	—
	Time slope variance	0.0032	—	—	—
	Covariance (Int,Time)	-0.040	—	—	—
	Residual SD (σ)	0.624	—	—	—

In the survival sub-model reported in Table 1, neither age nor sex has a statistically significant effect on survival time ($p > 0.05$). However, the association parameter (γ) is negative and highly significant ($p = 0.0001$), confirming that higher Hb levels are associated with a reduced risk of death.

Table 2: Performance criteria for the Joint Model (EM algorithm).

Criterion	AIC	RMSE
Joint Model (EM)	5119.95	0.488

As indicated in Table 2, the Akaike Information Criterion (AIC = 5119.95) reflects a good balance between model fit and complexity, while the relatively low RMSE value (0.488) demonstrates good agreement between the observed and predicted values, confirming the efficiency of the EM-based joint model.

6 CONCLUSIONS

The results of this study demonstrate that the joint model for longitudinal and survival data provides a reliable and effective framework for analyzing the relationship between time-varying biomarkers and survival probabilities in children with leukemia. The longitudinal analysis showed a gradual decline in hemoglobin (Hb) levels over time, reflecting disease progression and treatment effects. Age was positively associated with Hb levels, while females had lower Hb values than males, possibly due to physiological or therapeutic differences.

In the survival analysis, neither age nor sex had a significant direct effect on survival time, suggesting that other clinical or genetic factors may play a stronger role. The association parameter (γ) showed a negative and highly significant relationship, indicating that lower Hb levels increase the risk of death, confirming the predictive power of longitudinal biomarkers in survival estimation.

Performance evaluation criteria demonstrated the efficiency of the Expectation–Maximization (EM) algorithm in providing accurate and stable parameter estimates, even with incomplete data. Overall, the joint model serves as a valuable statistical tool that supports dynamic disease monitoring, enhances prediction accuracy, and aids in evidence-based medical decisions to improve outcomes for children with leukemia.

REFERENCES

[1] M.-H. Chen, J. G. Ibrahim, and D. Sinha, “A new joint model for longitudinal and survival data with a cure fraction,” Journal of Multivariate Analysis, vol. 91, no. 1, pp. 18-34, 2004.

[2] D. Ibrahim, L. Chu, and M.-H. Chen, “Basic concepts and methods for joint models of longitudinal and survival data,” Journal of Clinical Oncology, vol. 28, no. 16, pp. 2796-2801, 2010.

[3] N. Al-Huniti, “Dynamic predictions of survival in non-small cell lung cancer using joint modeling,” American Society for Clinical Pharmacology and Therapeutics (ASCPT) Annual Meeting, 2018.

[4] L. Wu, Y. Liu, G. Yi, and X. Huang, “Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues,” Journal of Probability and Statistics, vol. 2012, Article ID 640153, 2012.

[5] J. Murray and P. Philipson, “A fast approximate EM algorithm for joint models of survival and multivariate longitudinal data,” Computational Statistics & Data Analysis, vol. 170, 107437, 2022.

[6] M. J. Crowther, K. R. Abrams, and P. C. Lambert, “Joint modeling of longitudinal and survival data,” Stata Journal, vol. 13, no. 1, pp. 165-184, 2013.

[7] G. Babykina and G. Marot, “Longitudinal data analysis in high dimension,” M2 Internship Report, University of Lille, 2022.

[8] V. Medina-Olivares, F. Lindgren, R. Calabrese, and J. Crook, “Joint model for longitudinal and spatio-temporal survival data,” European Journal of Operational Research, vol. 327, no. 3, pp. 892-904, 2025.

[9] D. Rizopoulos, Notes on Joint Models for Longitudinal and Survival Data, Erasmus Medical Center, 2021.

[10] M. S. Shahrokhabadi, D.-G. Chen, S. J. Mirkamali, A. Kazemnejad, and F. Zayeri, “Marginalized two-part joint modeling of longitudinal semi-continuous responses and survival data with application to medical costs,” Mathematics, vol. 9, no. 20, pp. 2603-2622, 2021.

[11] T. H. Nguyen, E. Babykina, and G. Marot, “High-dimensional multivariate longitudinal data for survival analysis of cardiovascular event prediction,” BMC Medical Research Methodology, vol. 22, 2022.

[12] V. T. Nguyen, Integrating Longitudinal Data for Enhanced Survival Analysis: Methods and Applications, PhD Thesis, Université Paris Cité, Paris, France, 2024.

[13] M. Alkhathami, Joint Modeling of Longitudinal and Survival Data, PhD Thesis, Carleton University, Ottawa, Canada, 2021.