MODELLING LONGITUDINAL AND SURVIVAL DATA WITH MULTIPLE BIOMARKERS

Olayiwola, O. M[1]., Fagbamigbe, A. F[2]., Igbalajobi, M.M[1]., Olayiwola, O.D[1]., and Olayiwola, T.A[3]

[1]Department of Statistics, College of Physical Sciences, Federal University of Agriculture Abeokuta, Nigeria.

[2]Department of Epidemiology and Medical Statistics, College of Medicine, University of Ibadan, Nigeria.

[1]Department of Biochemistry, College of Biosciences, Federal University of Agriculture Abeokuta, Nigeria.

Corresponding author: olayiwolaom@funaab.edu.ng

Abstract

Background: Existing joint model for longitudinal and survival data captured both types of data, but there is dearth of information about methodologies that captured simultaneously the trajectories of multiple biomarkers over time. This study developed a joint statistical model that captured concurrent trajectories of multiple biomarkers using longitudinal and survival data. Data from the Mayo Clinic trial on Primary Biliary Cirrhosis was used to validate the model. The dataset comprised 424 patients that met eligibility criteria, with 312 actively participated in the trial. An additional 112 cases that participated not in the trial consented to basic measurements and survival monitoring

Methods: The joint model was developed by integrating a longitudinal sub-model (longitudinal outcomes over time) and a survival sub-model (the time until a specified event occurs) and compared with Mayor's models. The longitudinal sub-model was represented by a linear mixed-effects model and the survival sub-model by the Cox proportional hazards model. The two sub models were connected using a shared random effect to capture the correlation between longitudinal trajectory and event risk. The model parameters were estimated using the Expectation-Maximization algorithm and diagnostic checks were carried out to validate the model.

Results: The results revealed consistent trends in serum bilirubin levels, significant differences in serum cholesterol between placebo and D-penicillamine groups, and gender-related disparities in survival outcomes. A 55% observed survival rate highlighted positive health outcomes, while an 8% incidence of liver transplants underscored the complexity of the targeted medical conditions. An even distribution of participants between interventions ensured a fair comparison, emphasizing the efficacy of D-penicillamine while acknowledging the challenging nature of the addressed health conditions. Gender-specific analyses showed significant associations, with females exhibiting a hazard of survival approximately 0.4913 times that of males.

Conclusion: The survival model identified significant associations between survival time and biochemical measurements with high predictive accuracy.

Keywords: Statistical Models, longitudinal data, survival data, biomarkers, clinical trial

1. Introduction

The development of joint model for longitudinal and survival data to identifying biomarkers with strong prognostic capabilities combines statistical methodology, medical research, and data analysis. This type of model helps in understanding of disease progression, treatment efficacy and patient outcomes (McHunu et al., 2020).

Longitudinal data are repeated measurement taken over time from the same subject to capture the evolution of variables within individuals' survival data, on the other hand, deals with the time until an event of interest occurs, (Lawson et al., 2014a). Combining these types of data in a joint model provides a comprehensive approach to studying how diseases spread and evolve within populations and using this understanding to make predictions about future trends and outcomes related to the disease (Zhang et al., 2012).

The identification of biomarkers with strong prognostic capabilities is important in modern medicine (Li et al., 2021). Biomarkers are measurable indicators that reflect normal or pathological biological processes and they play a pivotal role in early disease detection, patient ratification, treatment monitoring and personalized medicine. Developing a sustainable joint model helps in identifying these biomarkers accurately. Survival analysis helps to assess and model the influence of socio-demographic, cultural, and economic factors on the age at first childbirth, (Fagbamigbe and Idemudia, 2016). Survival analysis concepts within a Bayesian framework is relevant for modelling the probability and progression of hidden diseases, (Olayiwola et al., 2020). Longitudinal measurements from the same subject can be correlated and survival times might vary due to patient-specific factors (Erango et al., 2018). The model must account for such correlations and heterogeneities. It may have missing values due to various reasons. Dealing with missing data appropriately is crucial to ensure accurate modelling (Huang et al., 2011).

Joint models are inherently complex because they need to capture both longitudinal trajectories and survival outcomes (Rustand et al., 2023). The development of a sustainable joint model involves a combination of statistical techniques, data preprocessing and validation steps. This includes handling missing data, scaling and transforming variables and identifying outliers. Identifying relevant biomarkers is crucial. Techniques like LASSO (Least Absolute Shrinkage and Selection Operator) can help select important features, (Niekerk et al., 2021). This joint model could involve shared random effects models, frailty models, or joint latent class models, among others. Bayesian or Likelihood-based methods are often employed for estimating model parameters. Markov Chain Monte Carlo (MCMC) techniques might be used for Bayesian estimation, (Law, 2002).

Cross-validation, bootstrap resampling, Concordance index (C-index), Time-Dependent Receiver Operating Characteristics (ROC) curve, brier score, calibration plots, external validation, model complexity and overfitting, clinical relevance and sensitivity analysis are used to assess the model's predictive performance and generalizability (Sweeting and Thompson, 2011). The model can predict patients' survival probabilities based on longitudinal measurements and biomarker values. It can assess the impact of treatments on disease progression and survival outcomes, (Alafchi et al., 2021). By analysing the estimated coefficients of biomarkers, the model helps identifying those with strong prognostic capabilities.

In the literature, there have been various efforts to model longitudinal and survival data simultaneously, The data occur together in medical and epidemiological studies, (Andrinopoulou et al., 2020; Baghfalaki et al., 2014; Köhler et al., 2018; Sweeting and Thompson, 2011). The proposed modelling approaches aim to capture the relationship between time-varying longitudinal measurements (such as biomarker levels) and the time until an event of interest. Some of the challenges and pitfalls of this model is misspecification, computational complexity, identification of shared latent variable that links the longitudinal and survival processes, incorrect results from the model due to assumption violation, selection bias in longitudinal data introduces bias into the survival analysis, and model validation may not fully capture the model's performance in predicting future events (Rustand et al., 2023).

There are certain challenges and limitations that affect sustainability of longitudinal and survival data modelling; incomplete or missing data and measurement errors, (Huang et al., 2011). Interpretation challenges, especially for non-experts can limit the utility of these models in clinical practice. .

(Henderson et al., 2000) used a semi-parametric approach in the joint modelling of Longitudinal and Time-to-Event Data. (Lawson et al., 2014b) work on joint analysis of time-to-event and multiple binary indicators of latent classes. (Alafchi et al., 2021), (Sun et al., 2019), (Henderson et al., 2000) revisited the likelihood Approach in Joint Modelling of Survival and Longitudinal Data using Semi-parametric Approaches with Time-Varying Coefficients in 2007 among others. Other authors who also studied the trend of repeated outcomes conditional on survival time include, (Martins et al., 2016). (Köhler et al., 2018; Law, 2002; Rustand et al., 2022; Sweeting and Thompson, 2011; Yu et al., 2004) presented joint modelling of longitudinal data and time to an event. Joint modelling with software application was presented by (Erango et al., 2018) and (Yuen and Mackinnon, 2016). According to the aforementioned literatures, homogeneity, normality, and other symmetric distribution assumptions are frequently utilized with a mixed effect model for the longitudinal component of the model. However, the majority of this research uses diverse samples whose measurements is highly skewed or contain some outliers.

(Baghfalaki et al., 2014; Huang et al., 2011; Rizopoulos et al., 2014a; Sène et al., 2014; Yuen and Mackinnon, 2016) discussed heterogonous random effects using a parameterization of the typical random effects. To study the impact of incorrectly specifying the random effects distributions on the parameter estimates and associated standard errors,(*Fully Exponential Laplace Approximations for the Joint Modelling of Survival and Longitudinal Data on JSTOR*, n.d.). This research study delved into developing a novel multivariate joint statistical model that handled both longitudinal and survival data with heterogeneous assumptions for the identification of biomarkers with strong prognostic capabilities.

2. Material and Method

2.1 Study design and setting.

This study introduced a joint statistical model to identify biomarkers with significant prognostic capabilities. Validation of the model was carried out using data from Mayo Clinic trial that investigated the effects of D-penicillamine compared to a placebo on individuals with Primary Biliary Cirrhosis.

2.2 Data Source and Study population

The dataset was obtained from Mayo Clinic trial that focused on Primary Biliary Cirrhosis (PBC) of the liver, (Dickson et al., 1989). Within the ten-year timeframe (1974 - 1984), a total of 424 PBC patients referred to the Mayo Clinic met the eligibility criteria for inclusion in a randomized placebo-controlled trial involving the drug D-penicillamine. Among these, the first 312 cases actively participated in the randomized trial, providing a dataset with predominantly complete information. Additionally, 112 cases from the same patient pool did not partake in the clinical trial but consented to recording basic measurements and agreed to be monitored for survival outcomes. Six of these cases were lost to follow-up shortly after diagnosis. Therefore, the dataset includes information on the remaining 106 cases who did not participate in the trial, along with the 312 individuals who were randomized participants.

2.3 Data collection

The Mayo Clinic trial on Primary Biliary Cirrhosis (PBC) of the liver employed a combination of research methodologies, including Randomized Controlled Trial (RCT) Procedures, administration of questionnaires and interviews, clinical assessments, laboratory tests, and follow-up with survival monitoring,

2.4 Ethical consideration

During the Mayo Clinic trial, participants willingly and knowingly gave their informed consent to take part in the trial

Study variables.

Case Number (A unique identifier for each participant in the study), Number of Days Between Registration and, Status (Categorized as 0 (alive), 1 (liver transplant), or 2 (dead)], Drug (Designated as 1 for D-penicillamine and 2 for placebo), Age in Days (The age of the participant measured in days), Sex (Coded as 0 for male and 1 for female), Serum Bilirubin (Measured in mg/dL), Serum Cholesterol (Measured in mg/dL), Albumin (Measured in gm/dL), Alkaline Phosphatase (Measured in U/liter), SGOT (Serum Glutamic Oxaloacetic Transaminase measured in U/ml), Platelets per Cubic mL / 1000 (The platelets count per cubic millilitre, scaled by a factor of 1000), Prothrombin Time (Measured in seconds) were considered.

2.5 Joint Model for Survival and Longitudinal Data

In joint modelling, a usual practice is to combine the linear mixed-effects sub-model (for the longitudinal process) and Weibull proportional hazards sub-model (for the time-to-event process). The survival sub-model includes the association parameter ϕ in the Weibull PH model, in which ϕu_i defines the nature of association structure between the two processes, ϕ measures the strength of association between the two processes. The joint model is fitted by the ML approach. Both the longitudinal and event processes were linked using an association parameter which can be a shared latent structure or shared random effect.

Rizopoulos et al proposed a joint model, where the main interest is in the time-to-event process, which is influenced by a longitudinal time-dependent covariate measured with error, (Rizopoulos et al., 2014a). Hickey developed a shared random effects model, where the focus is on both survival and longitudinal processes (Hickey et al., 2018).

This work considered the joint approach proposed by Rizopoulos et al (Rizopoulos et al., 2014a) with the heterogeneous assumption of random effect proposed by Baghfalaki et al (Baghfalaki et al., 2017).

Let $Y_i, i = 1, 2, 3, \ldots, n$ denote the vector of n_i longitudinal measurement for the ith individual such that $Y_i = \{y_i(t_{ij}), j = 1,2, \ldots, n_i)\}$ where $y_i(t_{ij})$ represent the longitudinal measurement for ith individual at time t_{ij}. If we have n subjects with their lifetimes represented by $T_1, T_2, \ldots, T_n$. If the data are right censored, T_i be the true event time and where $C_i > 0$ indicates a potential censoring time then, $t_i = \min(T_i, C_i)$ indicates the observed survival time for any ith individual, $i = 1, 2, 3, \ldots, n$. If δ_i represents the censoring indicator which is 0 for right-censored and 1 for completely observed individuals.

such that

$$\delta_i = \begin{cases} 1 & if \quad I(T_i \leq C_i) \\ 0 & if \quad otherwise \end{cases} \tag{1}$$

For each subject i, a pair of data will be observed for the survival outcome i.e.$\{T_i, \delta_i\}$, for $i = 1,2, \ldots, n$ and T_i is the true survival time according to whether the value of $\delta_i = 1 \; or \; 0$ respectively; .

If an event occurs at a time T_i, then after that event longitudinal measurements cannot be observed. Therefore, Y_i can be portioned into $Y_{i,observed} = m_i(t) = \{Y_i(t_{ij}): t_{ij} < T_i, j = 1,2, \ldots, n_i\}$ which contains all observed longitudinal measurements for the ith individual before the observed event time T_i, and $Y_{i,unobserved} = \{Y_i(t_{ij}): t_{ij} \geq T_i, j = 1,2, \ldots, n_i\}$which contains the longitudinal measurements that should have been taken until the end of the study. In this context, some individuals are missing (dropout or death). If the probability of the event occurring is dependent on the unobserved outcome, then missingness is non-ignorable or dropout is non-random.

Also, let $u_i = b_i$ (vector of random effects shared by both processes) with density function $f(b_i; \theta_b)$. If conditional independence of y_i and (T_i, δ_i) provided b_i, the joint conditional distribution is given by $f(y_i, T_i, \delta_i | b_i; \theta) = f(y_i, b_i; \theta_y) \times f(T_i, \delta_i | b_i; \theta_t)$ where $\theta = (\theta'_y, \theta'_t, \theta'_b)'$

2.6 Longitudinal Sub-model

The hazard function $h_i(t)$ in the hazard model depends on the true unobserved value of the longitudinal outcome, $m_i(t)$ at time t, Wu and Liu, (2012). However, the longitudinal measurements y_{ij} are collected with error on each subject at times t_{ij} ;i = 1, 2, · · · , N, j = 1, 2, · · · , n_i. Therefore, estimation of $m_i(t)$ and reconstruction of the true longitudinal history $M_i(t)$ is need for each subject to measure the impact of the longitudinal outcome to the hazard for an event, (Rizopoulos et al., 2014b). A suitable mixed-effects model with the observed repeated measurements $y_{ij} = \{y_i(t_{ij}), j = 1, 2, \cdots, n_i\}$ of each subject i, was fitted to describe the subject-

specific time evolutions. Sub-model for the longitudinal process is a linear mixed-effects (LME) model, given by

$y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + u_{i1} z_{1ij} + \cdots + u_{iq} z_{qij} + \epsilon_{ij},\ u_{ik} \sim N(0, \sigma_k^2),\ \ \mathrm{Cov}(u_{ik}, u_{ik})$ where $y_i = (y_{i1}, y_{i2}, \ldots, y_{ini})'$ is an n_i-dimensional vector of longitudinal responses for subject i, β is a p-dimensional vector for fixed effects parameters, X_i is an $(n_i \times p)$ design matrix of fixed effects β and u_i is a q-dimensional vector of random effects parameters that represents the characterization of between-individual variation. Also, Z_i is a $(n_i \times q)$ design matrix of random effects u_i, and ε_i is an n_i-dimensional vector of random errors that represents the characterization of within-individual variation.It is assumed that random effects u_i and random error ε_i follow a multivariate normal distribution with mean 0 and variance-covariance matrix D, within-individual error ε_i is independent, and u_i is assumed to be independent of the random errors ε_i (Nguyen et al., 2023).

The linear mixed effect model is a hierarchical two stage model because it allows the analysis of within-subject and between-subject sources of variation, where stage 1 specifies the within-subject variation, which is given by equation 2

$$Y_i = X_i\beta + Z_i U_i + \epsilon_i; i = 1,2, \ldots, N \tag{2}$$

and stage 2 specifies the between-subject variation, given by equation 3.

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_i(t), \\ m_i(t) = x_i'(t)\beta + z_i'(t)u_i, \\ u_i \sim N(0, D), \varepsilon_i(t) \sim N(0, \sigma_\epsilon^2), \end{cases} \tag{3}$$

where $z_i(t)$ is the design vector for the random effects $u_i \sim N(0, D)$ and $\varepsilon_i \sim N(0, R_i)$

As a result, we have equation 4.

$$E\begin{bmatrix} u_i \\ e_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{4}$$

The random effects u_i is the deviations of individual i from the population mean, while the mean parameters β is interpreted as the same as in a linear regression model.

R_i represents the diagonal matrix $\sigma_\epsilon^2 I_{ni}$, with I_{ni} being an $n_i \times n_i$ identity matrix.

The marginal mean vector and covariance matrix of the response vector $y_i(t)$ are

$$E(Y_i) = \mu_i = X_i\beta. \tag{5}$$

$$Var(Y_i) = V_i = Z_i D Z'_i + \sigma_\epsilon^2 I_{ni}. \tag{6}$$ The covariance of Y_i is given as

$$Var\begin{bmatrix} u_i \\ e_1 \end{bmatrix} = \begin{bmatrix} D & 0 \\ 0 & R_i \end{bmatrix} \tag{7}$$

And the conditional and marginal distributions of the response Y_i, is given as $Y_i | u_i \sim N(X_i\beta + Z_i U_i, \sigma_n^2 I_{ni})$.

2.7 Maximum Likelihood Estimation of Linear Mixed Effects Model parameters

The statistical inference of an LME model is based on the maximum likelihood and restricted maximum likelihood methods (Rizopoulos et al., 2014a).

Let $\theta = (\beta, \sigma_{ij}^2, D)$ denotes all parameters in LME model, $y_i = X_i\beta + Z_iU_i + e_i\ ;\ i = 1,2,\dots,N$ such that $U_i \sim N(0, D)$ and $e_i \sim N(0, R_i)$

The likelihood function for the observed data $y = (y_1' \dots y_N')'$ is given by

$$L(\theta|y) = \prod_{i=1}^{N} f(y_i : \theta) \quad (8)$$

$$= \prod_{i=1}^{N} f(y_i|u_i; \beta, \sigma_{ij}^2)\, f(u_i|D)du_i \quad (9)$$

where $f(y_i|u_i; \beta, \sigma_e^2)$ is the normal density with mean vector $(X_i\beta + Z_iu_i)$ and covariance matrix $\sigma_\varepsilon^2 I_{ni}$, and $f(u_i|D)$ is the normal density with mean vector 0 and covariance matrix D

Maximization of the likelihood function is based on an iterative algorithm, e{Expectation maximization (EM) algorithm or the Newton-Raphson method}. If the longitudinal responses of subjects are independent conditionally on their random effects, the log-likelihood of the LME model is given by

$$\rho(\theta) = \sum_{i=1}^{N} \log f(y_i|u_i; \beta, \sigma_{ij}^2)\ f(u_i|D)du_i \quad (10)$$

For a known covariance matrix V_i, the MLE of the fixed effects parameter β is given by

$$\hat{\beta} = (\sum X_i' V_i^{-\prime} X_i)^{-1} \sum X_i' V_i^{-\prime} y_i \quad (11)$$

which is an unbiased estimate of β. The asymptotic distribution of $\hat{\beta}$ is multivariate normal with the mean being the true value of β and the covariance given by

$$Cov(\hat{\beta}) = (\sum_{i=1}^{N} X_i'\, V_i^{-\prime}\, X_i)^{-1} \quad (12)$$

In case V_i is unknown, we use the estimate $\widehat{V_i}$ to find an estimate of β.

2.8 Survival Sub-model

Survival data are generally described and modelled in terms of survival function and the hazard function. Given the hazard function for the i^{th} individual as shown in equation $h(t_i|x_i, z_i, b_i) = h_0(t_i)\exp\{x'_i\beta + z'_i b_i\}$ (13)

the density function of survival time was expressed as

$$h^{\delta_i}(t_i|x_i, z_i, b_i) \times \exp\{-H_0(t_i)\exp\{x'_i\beta + z'_i b_i\}\}, \quad (14)$$

where $H_0(t) = \int_0^t h_0(u)du$, we have the p_2 and q_2 dimensional vectors of explanatory variables represented in the equation by x and z respectively.

$\beta = (\beta_1, \dots, \beta_{p_2})^T$ is a vector of p_2 dimension of time to event fixed-effect parameters; $b_i = (b_{i1}, \dots, b_{iq_2})^T$, is a vector of the time-to-event random effect of q_2 dimension such that $b_i \backsim N_{q2}(0, D_2)$.

The proposed model is built with the assumption that $b_i \sim \sum_{k=1}^{g} \pi_\kappa \mu_q(\mu_\kappa, D)$ and

$\epsilon_i \sim SN_{ni}\left(\sqrt{\frac{2}{\pi}}\ \delta_e, \Psi, \Delta_e\right)$, considering the population as heterogeneous.

Survival data are generally described and modelled in terms of two related functions: the survivor function and the hazard function. However, the natural approach is to postulate a proportional hazards model of the AFT form to describe the event hazard process at time t, as

$$h_i(m_i(t), w_i) = \left\{\lim_{\delta \to 0} P[t \le T_i < t + \delta_t | T_i \ge t, m_i(t), w_i]\right\}$$

Hazard function= $\lambda_0(g_i(t))\exp\ [w_i'\psi + \varphi m_i(t)],\ t > 0$ (15)

Where $M_i(t) = \{m_i(t), 0 \le s < t\ \}$ indicates the history of the true unobserved values of longitudinal covariate up to time point t,

$$g_i(t) = H_0(t) = \int_0^t exp\left(w_i'\psi + \varphi m_i(u)\right)du. \quad (16)$$

$\lambda_0(t)$ is the baseline hazard function, and w_i is a vector of baseline covariates with a corresponding vector of regression coefficients, ψ. The parameter ϕ measures the impact of the underlying longitudinal outcome on the risk for an event.

The corresponding survival function depends on the whole history of the true unobserved longitudinal process up to time point t, the $M_i(t)$. That is

$$S_i(t|m_i(t), w_i) = P(T_i > t, m_i(t), w_i)$$

$$S_i(t) = \lambda_0\left(g_i(t)\right) \quad (17)$$

$$= exp\left[-\int_0^t \lambda_0(t) \exp[w_i'\psi + \varphi m_i(t)ds]\right] \quad (18)$$

Both the hazard function and survivor functions are written as a function of the baseline hazard λ_0(t). To avoid the issue of underestimation of the standard errors of the parameter estimates in the joint model, the Weibull was used.

In this study, Weibull model was studied for $\lambda_0(t)$. The survival times follow a Weibull distribution,$W(\lambda, \tau)$, with the scale parameter λ and shape parameter τ. The (Weibull model) hazard function was obtained as equation (19)

$$h(t) = \lambda\tau t^{\tau-1} \quad , 0 \le t < \infty. \quad (19)$$

If τ = 1, the survival times with exponential distribution as a special case of the Weibull distribution. For other values of τ, the hazard function increases or decreases monotonously for τ > 1 and τ < 1, respectively.

For this choice of hazard function with the distribution $W(\lambda \exp(w_i'\gamma)\, \tau)$, the (Weibull model) survival sub-model was written as equation (20)

$$h_i(m_i(t), w_i) = h_0(t) \exp[w_i'\psi + \varphi m_i(t)] \quad (20)$$

$$= \lambda\tau t^{\tau-1} \exp[w_i'\psi + \varphi m_i(t)] \quad (21)$$

$$= \lambda\tau t^{\tau-1} \exp[w_i'\psi + \varphi m_i(t)],\ (\lambda = e^{\lambda_0}) \quad (22)$$

where $m_i(t)$ involves random effects u_i, which is independent N(0, D), with covariance D = D(ξ) (Crowther et al., 2013).

Assume a data set given as $\{(t_i, \delta_i); i = 1, 2, \ldots, N\}$, the likelihood function for the frailty model parameters is obtained as

$$L(\psi, T, D, \varphi) = \prod_{i=1}^{N} \int_{-\infty}^{\infty} [f(t_i|u_i;\ \psi, T, \varphi)]^{\delta_i} [s(t_i|u_i;\ \psi_i, T, \varphi)]^{1-\delta_i} f(u_i|D) du_i \quad (23)$$

$$= \prod_{i=1}^{N} \int_{-\infty}^{\infty} [h(t_i|u_i;\ \psi_i, T, \varphi)]^{\delta_i} s(t_i|u_i;\ \psi_i, T, \varphi)\ f(u_i|D) du_i \quad (24)$$

where $f(t_i|u_i;\ \gamma,\ \tau,\ \phi)$ is the conditional density function of the event time, given the frailty u_i, $S(t_i|u_i;\ \gamma,\ \tau,\ \phi)$ is the conditional survivor function for the i th subject at time t_i, and $f(u_i|D)$ is the conditional density of the random effects u_i. In this setting, the density function for the random effect u_i is given by $u_i \sim \sum_{k=1}^{g} \pi_\kappa N_q\ (\mu_\kappa,\ D)$ instead, as expressed in equation (25).

$$f(u_i|D) = \frac{\exp[-1/2 U_i^{-1} D^{-1} U_i]}{(2\pi)^{\varepsilon/2} |D|^{1/2}} \quad (25)$$

whereas the conditional density for survival times is given by the Weibull distribution

$$f(t_i|u_i;\ \psi, T, \varphi) = [h(t_i|u_i;\ \psi, T, \varphi)]^{\delta_i} s_i(t_i|u_i;\ \psi, T, \varphi) \quad (26) = \left[T t^{T-1} \exp\left(\lambda w_i' \psi + \varphi m_i(t)\right)\right]^{\delta_i} \exp\left[-T t^{T-1} \exp\left(\lambda w_i' \psi + \varphi m_i(t)\right)\right] \quad (27)$$

The corresponding log-likelihood function for the frailty sub-model is given by

$$\rho(\psi, T, D, \varphi) = \sum_{i=1}^{N} \int_{-\infty}^{\infty} [\delta_i \log h_i\ (t_i|u_i;\ \psi_i, T, \varphi) + \log s_i(t_i|u_i;\ \psi, T, \varphi)]\ f(u_i|D) du_i$$

2.9 Estimates of model parameters

Let θ = (γ, τ,D, ϕ) denote the vector of all model parameters that need to be estimated. The ML estimators of θ may be obtained by using a numerical maximization method. Given some initial estimates $\theta^{(0)}$, we can obtain the ML estimates by solving the Newton-Raphson iterative equations

$\theta^{(m+1)} = \theta^{(m)} + \{I(\theta^{(m)})\}-1\ \{U(\theta^{(m)})\}$

form = 0, 1, 2, · · · , where $U(\theta^{(m)})$ is the likelihood score function $U^{(\theta)}$, given by

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^{N} \int_{\infty}^{-\infty} \left[\delta_i \frac{\partial \log h_i(t_i|u_i;\ \psi_i, T, \varphi)}{\partial \theta} \mid \frac{\partial \log S_i(t_i|u_i;\ \psi, T, \varphi)}{\partial \theta}\right] f(u_i|t_i; \theta) du_i,$$

When evaluated at $\theta^{(m)}$, the Fisher information matrix $I(\theta)^{(m)}$ will be obtained from the first derivative of the score function U(θ) with respect to θ evaluated at $\theta^{(m)}$.

$$I(\theta) = \sum_{i=1}^{N} \int_{\infty}^{-\infty} \left[\delta_i \frac{\partial^2 \log h_i(t_i|u_i;\ \psi_i, T, \varphi)}{\partial \theta \theta'} + \frac{\partial^2 \log S_i(t_i|u_i;\ \psi, T, \varphi)}{\partial \theta \theta'}\right] f(u_i|t_i; \theta) du_i + \sum_{i=1}^{N} \int_{\infty}^{-\infty} \left[\delta_i \frac{\partial \log h_i(t_i|u_i;\ \psi_i, T, \varphi)}{\partial \theta} + \frac{\partial \log S_i(t_i|u_i;\ \psi, T, \varphi)}{\partial \theta}\right]$$

$$\times \left[\delta_i \frac{\partial \log h_i(t_i|u_i;\ \psi_i, T, \varphi)}{\partial \theta'} + \frac{\partial \log S_i(t_i|u_i;\ \psi, T, \varphi)}{\partial \theta'}\right] f(u_i|t_i;\theta)du_i$$

$$-\sum_{i=1}^{N}\int_{\infty}^{-\infty}\left[\delta_i \frac{\partial \log h_i(t_i|u_i;\ \psi_i, T, \varphi)}{\partial \theta} + \frac{\partial \log S_i(t_i|u_i;\ \psi, T, \varphi)}{\partial \theta}\right] f(u_i|t_i;\theta)du_i$$

$$\times \int_{\infty}^{-\infty}\left[\delta_i \frac{\partial \log h_i(t_i|u_i;\psi_i,T,\varphi)}{\partial \theta\prime} \mid \frac{\partial \log S_i(t_i|u_i;\psi,T,\varphi)}{\partial \theta\prime}\right] f(u_i|t_i;\theta)du_i. \quad (28)$$

The likelihood function does not have a closed form. Also, both the likelihood score function and Fisher information matrix involve the calculation of multi-dimensional integrations with respect to the conditional distribution of the random effects, u_i given time t_i i.e $u_i|t_i$, which does not have a closed form. The integrations involving the conditional expectations may be computed numerically using existing software (Rizopoulos, 2012). In this study, R function was used to obtain the maximum likelihood (ML) estimates $\hat{\theta}$ of θ in the Weibull frailty model.

2.10 The Proposed Joint Model

Longitudinal Part of the model

$y_i = X_i\beta + Z_i u_i + \mathcal{E}_i$

Where y_i is the longitudinal response for subject i, X_i is the design matrix for the fixed effects in the longitudinal sub-model, β is the vector of fixed effects coefficients, Z_i is the design matrix for the random effects in the longitudinal sub-model, u_i is the vector of random effects for subject i and $\mathcal{E}_i$ is the error term for subject i.

Survival part of the model

$y_i = X_i\beta + Z_i u_i + \mathcal{E}_i$ and $h_i(t) = \lambda T t^{T-1}\exp\{\varpi_i{}^T\psi\}$

Where:

$h_i(t)$ is the hazard function for an individual with covariates x at time t, λ is the scale parameter, γ is the shape parameter, ψ is a vector of regression coefficients and ϖ is a vector of covariates for an individual.

The joint model was linked through shared random effects by allowing the random effects u_i from the longitudinal model to influence the baseline hazard function or survival model in some way.

The Link Structure

For implementation, $V_i(t) = \varrho U_i(t)$ was used to capture the dependence between the longitudinal and time-to-occurrence of an event sub-models; where ϱ is measure of link or association induced by the fitted longitudinal responses. Random intercept and random slopes can equally be used to formulate the association structure (Henderson et al., 2000).

Log-Logistic Sub-Model

By using equation $h_i(t) = h_0\big(g_i(t)\big)\exp(\boldsymbol{z'_i\beta} + V_i(t))$ and $V_i(t) = \varrho U_i(t) = \varrho \boldsymbol{r'}_i(t)\boldsymbol{b}_i$, the hazard function at time t_i can be written as

$$h_i(t_i) = \frac{k\rho\,(\rho g_i(t_i))^{k-1}}{1 + (\rho g_i(t_i))^k} \exp(z'_i\beta + \varrho r'_i(t_i)b_i), \quad (29)$$

where $g_i(t_i) = \int_0^{t_i} \exp(z'_i\beta + \varrho r'_i(u)b_i)du$ and

$h_o(g_i(t_i)) = k\rho\,(\rho g_i(t_i))^{k-1} / [1 + (\rho g_i(t_i))^k]$.

Also, the survival function is given as:

$$S_i(t_i) = S_o\big(g_i(t_i)\big) = [1 + (\rho g_i(t_i))^k]^{-1} \quad (30)$$

The density function of (t_i, δ_i) given b_i and θ can be determined using equations (29) and (30) as

$$f(t_i, \delta_i | b_i, \theta) = \{\frac{k\rho\,(\rho g_i(t_i))^{k-1}}{1+(\rho g_i(t_i))^k} \exp(z'_i\beta + \varrho r'_i(t_i)b_i)\}^{\delta_i} \times [1 + (\rho g_i(t_i))^k]^{-1}$$

(31)

Weibull Sub-Model

Weibull model (or a Cox proportional hazard model) may be linked to the longitudinal model through shared random effects.

$T_i \sim$ Weibull $(z'_i\beta + \varrho r'_i(t_i)b_i)$ (32)

Where $\beta = \big(\beta_1, \beta_2, \dots, \beta_p\big)'$, is a p −dimensional vector of fixed parameters, $z_i = \big(z_{i1}, z_{i2}, \dots, z_{ip2}\big)'$ is a p_2- dimensional vector of explanatory variables, b_i is shared between the two models and r is a q- dimensional vector of association parameters. If $r = 0$, the event time and the longitudinal measurements are independent. Also, the scalar ϱ is the shape parameter

To obtain the hazard function at time t_i, we can write $E(Y_i) = \mu_i = \mu_0 + \sqrt{\frac{2}{\pi}}\,\delta^2$ $and\ Var(Y_i) = V_i = \Psi + \left(1 - \frac{2}{\pi}\right)\Delta^2$. and $V_i(t) = \varrho U_i(t) = \varrho \boldsymbol{r}'_i(t)\boldsymbol{b}_i$ as

$$h_i(t_i | z_i, \boldsymbol{r}, \boldsymbol{b}_i) = h_{0i}(t_i)\exp(\boldsymbol{z}'_i\boldsymbol{\beta} + \varrho \boldsymbol{r}'_i(t_i)\boldsymbol{b}_i), \quad (33)$$

$= k\rho\,(\rho g_i(t_i))^{k-1}\exp(\boldsymbol{z}'_i\boldsymbol{\beta} + \varrho \boldsymbol{r}'_i(t_i)\boldsymbol{b}_i)$

where

$g_i(t_i) = \int_0^{t_i} \exp(\boldsymbol{z}'_i\boldsymbol{\beta} + \varrho \boldsymbol{r}'_i(u)\boldsymbol{b}_i)du$ and $h_o(g_i(t_i)) = k\rho\,(\rho g_i(t_i))^{k-1}$

where $h_{0i}(t_i)$ is the base line hazard function. The baseline hazard was assumed to be a step function,

$h_0(t) = h_k$, for $s_{k-1} < t \le s_k, k = 1, 2, \dots, K$ where $0 = s_0 < s_1 < s_2 < \dots < s_k < \infty$ is a partition of $(0, \infty)$ and K indicates the number of steps for the baseline hazard. Hence, the cumulative baseline hazard is given by

$H_0(t) = \left(h_j\left(t - s_{j-1}\right) + \sum_{i=1}^{j-1} h_i(s_i - s_{i-1})\right) I(t \in (s_{i-1}, s_j))$.

Sensitivity analysis of the results with respect to different values of K can be investigated. Assuming $h_0\ (t_i) = r{t_i}^{r-1}$ the proportional hazard model reduced to Weibull model. In application section, Weibull and the Cox models was used for analyzing the dataset.

In the above-mentioned structures, typical models assume that the random effects b_i follows a multivariate normal distribution with mean 0 and covariance matrix D. This model is sometimes called the homogeneity mixed model. In contrast, the heterogeneity model was introduced.

Hence, the survival function can be given as

$$S_i(t_i) = S_o\left(g_i(t_i)\right) = \exp[-(\rho g_i(t_i))^k]. \quad (34)$$

Also, both equation (33) and (34) can be used to express the density function (t_i, δ_i) given b_i and θ under Weibull model as

$$f(t_i, \delta_i | \boldsymbol{b_i}, \boldsymbol{\theta}) = \{k\rho\left(\rho g_i(t_i)\right)^{k-1} \exp(\boldsymbol{z'}_i\boldsymbol{\beta} + \varrho \boldsymbol{r'}_i(t_i)\boldsymbol{b}_i)\}^{\delta_i} \times \exp[-(\rho g_i(t_i))^k], \quad (35)$$

The heterogeneity model

The proposed model is based on the following assumptions:

$$u_i \sim \sum_{k=1}^{g} \pi_\kappa N_q(\mu_\kappa,\ D)$$

Where g is the number of components such that the probability of belonging to component k is π_κ and $\sum_{k=1}^{g} \pi_k = 1$. Also, μ_κ is the mean of the kth component and each component has the same covariance matrix D.

Further, $E[\ \boldsymbol{u_i}] = \sum_{k=1}^{g} \pi_k \mu_k$, and var $[[\boldsymbol{u_i}] = \sum_{k=1}^{g} {\mu'}_k \pi_k \mu_k (1 - \pi_k) + \boldsymbol{D}$.

$$\epsilon_i \sim SN_{ni}\left(\sqrt{\frac{2}{\pi}}\ \boldsymbol{\delta}_e, \boldsymbol{\Psi},\ \boldsymbol{\Delta}_e\right) .$$

where, $y_i(t) \sim SN_{n,i}(\mu_0, \boldsymbol{\Psi}, \boldsymbol{\Delta}_e)$,$\mu_0 \in \mathbb{R}^n$ is a location vector, $\boldsymbol{\Psi}$ is a scale matrix (n x n positive definite matrix), $\boldsymbol{\Delta}_e$ is the skewness matrix (n x k); if $\boldsymbol{\Delta}_e$ is set at 0 then, we have the usual symmetric multivariate normal distribution. $\boldsymbol{\delta} = (\delta_1, \delta_2 \ldots \delta_n)^T$ is the skewness parameter vector.

where β denotes the vector of the regression coefficients for the fixed effects covariates $x_1$and z_i denotes the covariate vector for the random effects $\boldsymbol{u_i}$. The fixed and random effects refer to the population-average and subject-specific effects, respectively. The error terms ${\varepsilon_i}^*(t)$ are mutually independent, skew normal distribution with variance σ_ϵ^2 , and independent of $\boldsymbol{u_i}$, (Rizopoulos, 2012). The random effects $\boldsymbol{u_i}$ in the model not only incorporate heterogeneity in the data but also incorporate correlation between the multiple measurements within each individual or cluster. The random effects $\boldsymbol{u_i}$ follow a heterogeneous normal distribution with covariance matrix D.

The random effects u_i is the deviations of individual i from the population mean, while the mean parameters β is regression coefficients

The proposed Joint Model assumes that the random effects $u_i \sim \sum_{k=1}^{g} \pi_\kappa N_q(\mu_\kappa, D)$ and the error term $\epsilon_i \sim SN_{ni}\left(\sqrt{\frac{2}{\pi}}\,\delta_e, \Psi, \Delta_e\right)$.

The survival model here may as well be written as

$$h_i(\mathcal{M}_i(t), \omega_i) = h_0(t)\exp\{\omega_i'\psi + \varphi[x_i'(t)\beta + z_i'(t)u_i]\}. \quad (36)$$

Under the Weibull PH model, $h(t) = \lambda\tau t^{\tau-1}$, $0 \le t < \infty$ the above hazard function may be written a $= \lambda\tau t^{\tau-1}\exp[w_i'\psi + \varphi m_i(t)]$, $(\lambda = e^{\lambda_0})$

$$h_i(\mathcal{M}_i(t), \omega_i) = \lambda\tau t^{\tau-1}\exp\{\omega_i'\psi + \varphi[x_i'(t)\beta + z_i'(t)u_i]\}, \left(\lambda = e^{\lambda_0}\right) \quad (37)$$

The joint model is given by

$$\begin{cases} y_i(t) = [x_i'(t)\beta + z_i'(t)u_i] \equiv m_i(t) + \varepsilon_i(t), \\ h_i(\mathcal{M}_i(t), \omega_i) = Tt^{T-1}\exp\{\omega_i'\psi + \varphi[x_i'(t)\beta + z_i'(t)u_i]\}, \\ u_i \sim \sum_{k=1}^{g} \pi_\kappa N_q(\mu_\kappa, D), \epsilon_i \sim SN_{ni}\left(\sqrt{\frac{2}{\pi}}\,\delta_e, \Psi, \Delta_e\right) \end{cases} \quad (38)$$

where in the longitudinal sub-model, $x_i(t)$ and $z_i(t)$ are vectors of possibly time-dependent covariates associated with the p-vector of fixed effects β and the q-vector of individual random effects u_i, with $u_i \sim \sum_{k=1}^{g} \pi_\kappa N_q(\mu_\kappa, D)$. The error terms $\varepsilon_i(t)$ and u_i are assumed independent. In the survival submodel, $h_0(t) = \lambda\tau t^{\tau-1}$ is the baseline hazard when survival times follow the Weibull distribution.

Usually, the baseline hazard is parametric (e.g., Weibull, piecewise constant, or a small number of B-splines). It is rare to keep $h_0(t)$ unspecified (like in the Cox model); the partial likelihood of the Cox model cannot be employed, and the full likelihood has to be defined. As a solution, one might consider a piecewise constant function with jumps at each event time, but this would produce too many parameters and lead to computational problems, (Hsieh et al., 2006). The vector w_i denotes baseline covariates associated with the vector of coefficients γ, while the multivariate function of marker $m_i(t) = x_i'(t)\beta + z_i'(t)u_i$ is associated with the parameter ϕ, which quantifies the degree of association between the longitudinal outcome evaluated at time t and the corresponding hazard for an event

Another commonly used joint model framework is to link the survival and longitudinal models via shared random effects, called shared parameter models. In this case, the random effects may be interpreted as a summary of individual-specific longitudinal characteristics, or a latent process (shared variables) which governs both longitudinal and event progressions. Such a shared parameter model may be written as

$$\begin{cases} y_i = X_i\beta + Z_i u_i + \mathcal{E}_i \\ h_i(t) = Tt^{T-1}\exp\{\omega_i'\psi + \varphi' u_i\}, \end{cases} \quad (38)$$

where the Weibull PH model and LME model share the same random effects u_i. This joint model framework is frequently used when the survival risk is influenced by summaries of the longitudinal process (e.g., individual-specific intercepts and slopes)

2.11 Estimation of Joint Model Parameters

The method of estimation of this joint model parameters follows the maximum likelihood (ML) method.

Given observed data $\{y_i, t_i, \delta_i\}$; (i = 1, 2, · · · , N) from both survival and longitudinal outcomes, the joint likelihood is given by

$$L(\theta|y) = \prod_{i=1}^{N} f\left(t_i, \delta_{i,} y_i \middle| \theta\right) = \prod_{i=1}^{N} \int f\left(t_i, \delta_{i,} y_i, u_i \middle| \theta\right) du_i =$$
$$\prod_{i=1}^{N} \int f(t_i|u_i;\ \psi, T, \varphi)\, f(y_i|u_i; \beta, \sigma_e^2) f(u_i|D) du_i \quad =$$
$$\prod_{i=1}^{N} \int f\ (t_i|u_i;\ \psi, T, \varphi) \prod_{j=1}^{n_i} f\ (y_i|u_i; \beta, \sigma_e^2)\ f(u_i|D) du_i \ \ldots\ldots\ldots\ldots\ldots\ldots \quad (39)$$

where θ denotes the vector of all model parameters that need to be estimated. The conditional density for the Weibull survival time t_i takes the form;

$$f(t_i|u_i;\ \psi, T, \varphi) = \{h_i(t_i|u_i;\ \psi, T, \varphi)\}^{\delta_i} S_i(t_i|u_i;\ \psi, T, \varphi)$$
$$= \left\{Tt^{T-1}\exp(\omega_i'\psi + \varphi'^{u_i})^{\delta_i} \exp(-t_i{}^{T-1}\exp\ (\omega_i'\psi + \varphi'^{u_i})\right\}.$$

The conditional density for the longitudinal outcome y_{ij} is having mean $\mu_{ij} = x_{ij}'\beta + z_{ij}'u_i$ and variance σ_ε^2 . The conditional density for of $y_i = (y_{i1}, y_{i2}, \ldots, y_{ini})'$ is given by

$$f(y_i|u_i; \beta, \sigma_\varepsilon^2) = \prod_{j=1}^{n_i} f\left(y_{ij}\middle|u_i; \beta, \sigma_\varepsilon^2\right)$$

$$= (2\pi\sigma_\varepsilon^2)^{-\frac{ni}{2}} \exp\left\{-1\middle|\middle|y_i - x_{ij}'\beta + z_{ij}'u_i\middle|\middle|^2/2\sigma_\varepsilon^2\right\},$$

where $\left|\left|x\right|\right| = \{\sum_i x_i{}^2\}^{1/2}$is the norm of the Euclidian vector. The two outcome processes are linked via the random effects b_i, $b_i \sim \sum_{k=1}^{g} \pi_\kappa N_q\ (\mu_\kappa,\ D)$. When the association parameter ϕ = 0, the joint analysis is equivalent to the separate analysis.

The observed data log-likelihood for all individuals in the study can be formulated as

$$\ell(\theta) = \sum_{i=1}^{N} \log \int f\left(t_i, \delta_{i,} y_i, b_i \middle| \theta\right) db_i,$$
$$= \sum_{i=1}^{N} \log \int f(t_i|b_i;\ \psi, T, \varphi) \prod_{j=1}^{n_i} f\left(y_{ij}\middle|b_i; \beta, \sigma_\varepsilon^2\right) f(b_i|D) db_i \quad (40)$$

The ML of estimator of θ was obtained by maximizing the log-likelihood function with respect to θ using a Newton-Raphson iterative algorithm.

2.12 Model Diagnostic Check and comparison.

Diagnostic checks were employed to evaluate the adequacy of fit for the proposed joint models and the fitted model was compared with existing model (Mayor Model), (Dickson et al., 1989). The same dataset and variables were used in both models.

3. Results and discussion

From the Kaplan-Meier estimate presented in Figure 1, the D-penicil group demonstrates slightly higher survival rates than the placebo group after one month of follow-up. However, this trend changed after six months, wherein the placebo group began to exhibit higher survival rates. The difference in survival between the two groups appeared to diminish by the ninth month, and by the end of fourteen months of follow-up, there is a noticeable decline in survival for the placebo group.

Figure 1: Probability of survival for the placebo and D-penicil treatment groups.

From the Cox proportional hazards model, the dependent variable, which is the outcome being modelled, is survival time (years) and censoring status (status2), where status2 indicates whether the event of interest (death, transplant, alive) occurred or not. The independent variables, also known as predictor variables or covariates, are drug, sex, serum bilirubin, serum cholesterol, albumin, alkaline, SGOT, platelets and prothrombin. These variables are assumed to influence the survival time and censoring status of individuals in the dataset.

In table 1, the coefficient for the variable "sex-female" is -0.711, indicating that females have a hazard of survival that is approximately 0.491 times that of males when other variables are held constant. The associated p-value (<0.001) is highly significant, and provides strong evidence that gender is indeed associated with survival in the studied population. Moreover, the Likelihood ratio test, Wald test (261.1), and Score (log-rank = 302) test all yielded extremely low p-values (p = < <0.001), underscoring the overall statistical significance of the model. This Cox proportional hazards model revealed significant associations between survival time and serum bilirubin, serum cholesterol, albumin, alkaline, platelets, and prothrombin, even after adjusting for other covariates. The collective statistical significance of the model suggested that at least one of the considered predictors plays a crucial role in influencing survival outcomes. Furthermore, the Concordance statistic, measuring predictive accuracy, was reasonably high at 0.757 (standard error = 0.013), indicating the model's effectiveness in predicting survival times.

Table1: Estimates of survival model

The dependent variables in the Linear Mixed-Effects Model as shown in Table 2 are the biochemical measurements (serum bilirubin, serum cholesterol, albumin, alkaline, platelets, prothrombin and SGOT). The independent variables are drug (this is a fixed effect, meaning it's a predictor variable that is of interest for assessing the relationship with the dependent variables), year (this is also a fixed effect, representing time) and the interaction term drug * year.

Table 2: Linear mixed-effects model estimates

In terms of fixed effects coefficients, as shown in Table 2, the model provided valuable insights into the expected values of serum bilirubin, serum cholesterol, albumin, alkaline, platelets, and prothrombin. The estimated intercept is 3.447, indicating the anticipated value of these

measurements when all predictors are set to zero. The coefficient for drug-D-penicillamine is -0.736, signifying the alteration in biochemical measurements (serum bilirubin, serum cholesterol, albumin, alkaline, platelets, and prothrombin) associated with the use of drug-D-penicillamine in comparison to the placebo. Meanwhile, the coefficient for the variable "year" is 0.896, representing the anticipated change in biochemical measurements for each additional year. Additionally, the interaction term between drug-D-penicillamine and year yields a coefficient of -0.165. This implied that the impact of the D-penicillamine on survival changes over time, and the rate of change is captured by the interaction term. The negative sign indicated a decreasing effect between the D-penicillamine and survival as time progressed.

The standard deviation of the random intercepts across various ID levels is calculated as 4.0173 as shown in table 2, offering insights into the variability in baseline biochemical measurements between different individuals. Similarly, the standard deviation of the random slopes for distinct ID levels is 1.111, captured the variability in the rate of change of biochemical measurements over time. The high correlation coefficient of 0.982 between random slopes and intercepts indicated a strong relationship, suggesting that individuals with higher baseline measurements tend to experience steeper changes over time.

The residual standard deviation, quantified at 1.910, Table 2, denoted the unaccounted variability in serum bilirubin, serum cholesterol, albumin, alkaline, platelets, and prothrombin after incorporating both fixed and random effects. This metric served as a measure of the model's ability to explain the observed variations in the data. The log-restricted-likelihood, standing at -2842.63 for the linear mixed-effects model, reflected the model's goodness of fit. This model, encompassing fixed effects associated with drug, year, and their interaction, as well as random effects accommodating individual-level variations, provided valuable insights into the temporal dynamics of serum bilirubin, serum cholesterol, albumin, alkaline, platelets, and prothrombin. The log-restricted-likelihood serves as a critical metric for model evaluation, and the consideration of random effects aids in capturing the nuanced individual-level variations in both baseline levels and temporal trajectories of biochemical measurements.

The findings derived from the joint model provided valuable insights into the intricate connections between the longitudinal and event processes. Within the longitudinal process, the estimated intercept stands at 0.554, accompanied by a standard error of 0.033, a z-value of 16.765, and a remarkably significant p-value of less than 0.0001. This intercept signified the expected values of serum bilirubin, serum cholesterol, albumin, alkaline, platelets, and prothrombin when all other predictors were at zero. The coefficient for the 'year' variable is 0.185, indicating the anticipated change in serum bilirubin, serum cholesterol, albumin, alkaline, platelets, and prothrombin for a one-unit increase in the 'year' variable. The coefficient for 'drug-D-penicillamine' is 0.031, suggesting a modest impact on serum bilirubin, serum cholesterol, albumin, alkaline, platelets, and prothrombin, although statistical significance is achieved ($p = 0.040$). The interaction term 'drug-D-penicillamine and year' has a coefficient of 0.013, signifying how the effect of 'drug-D-penicillamine' changes for each additional year, this interaction is statistically significant ($p = 0.016$). Moving to the event process, the coefficient for 'drug-D-penicillamine' is 0.048, accompanied by a standard error of 0.182, a z-value of 0.263, and a p-value of 0.039. This suggests that the impact of 'drug-D-penicillamine' on the event process is statistically significant. The coefficient for 'sex-female' is 0.324, indicating that females exhibit a high hazard compared to males. This effect is statistically significant ($p = 0.016$). On the other hand, the coefficient for association term or the association parameter between the longitudinal and survival processes is

1.258, with a standard error of 0.096, a z-value of 13.175, and a highly significant p-value of less than 0.0001, suggesting a substantial impact on the event process. Additionally, spline terms 'bs1' to 'bs9' contribute to the spline-approximated baseline risk function.

The proposed joint model, as evidenced by a log-likelihood of 31485.34, Information Criterion (BIC) of -64226.81, Integrated Completed Likelihood (ICL) of -64482.31 and coefficient of determination (0.891), Table 3, demonstrates a superior fit compared to the existing model (Mayor Model). The higher log-likelihood indicates that the proposed model better explains the data. Moreover, the lower BIC and ICL values suggest that the proposed model more effectively captures underlying patterns in serum bilirubin, serum cholesterol, albumin, alkaline, platelets, and prothrombin and higher coefficient of determination compared with the existing model showed that it has a better goodness of fit.

Table 3: Estimates for the proposed and existing Models

4. Conclusion

The proposed joint model, which integrates both longitudinal and survival data, enhances predictive accuracy compared to using survival or longitudinal models independently. It provides a more comprehensive understanding of how changes in longitudinal measurements influence the timing and probability of survival events, enabling better-informed clinical decision-making and patient care.

4.1 Recommendations

The joint model offered a powerful framework for capturing the intricate interplay between longitudinal processes and survival outcomes. By simultaneously modelling longitudinal trajectories and time-to-event data, this model provided a nuanced understanding of how changes in longitudinal measurements influence the risk of experiencing events such as disease progression, relapse, or mortality. This comprehensive approach enables researchers and clinicians to gain deeper insights into the underlying mechanisms driving disease progression or treatment response, facilitating more informed decision-making in clinical practice and research settings.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial or personal interests that could have appeared to influence the work reported in this paper.

REFRENCE

Alafchi, B., Mahjub, H., Tapak, L., Roshanaei, G., and Amirzargar, M. A. (2021). Two-Stage Joint Model for Multivariate Longitudinal and Multistate Processes, with Application to Renal Transplantation Data. *Journal of Probability and Statistics*, *2021*(1), 6641602. https://doi.org/10.1155/2021/6641602

Andrinopoulou, E. R., Nasserinejad, K., Szczesniak, R., and Rizopoulos, D. (2020). Integrating latent classes in the Bayesian shared parameter joint model of longitudinal and survival outcomes. *Statistical Methods in Medical Research*, *29*(11), 3294–3307. https://doi.org/10.1177/0962280220924680/SUPPL_FILE/SJ-PDF-1-SMM-10.1177_0962280220924680.PDF

Baghfalaki, T., Ganjali, M., and Hashemi, R. (2014). Bayesian Joint Modeling of Longitudinal Measurements and Time-to-Event Data Using Robust Distributions. *Journal of Biopharmaceutical Statistics*, *24*(4), 834–855. https://doi.org/10.1080/10543406.2014.903657

Baghfalaki, T., Ganjali, M., and Verbeke, G. (2017). A shared parameter model of longitudinal measurements and survival time with heterogeneous random-effects distribution. *Journal of Applied Statistics*, *44*(15), 2813–2836. https://doi.org/10.1080/02664763.2016.1266309

Crowther, M. J., Abrams, K. R., and Lambert, P. C. (2013). Joint Modeling of Longitudinal and Survival Data. *The Stata Journal*, *13*(1), 165–184. https://doi.org/10.1177/1536867X1301300112

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, *10*(1), 1–7. https://doi.org/10.1002/HEP.1840100102

Erango, M. A., Goshu, A. T., Erango, M. A., and Goshu, A. T. (2018). Bayesian Joint Modelling of Survival Time and Longitudinal CD4 Cell Counts Using Accelerated Failure Time and Generalized Error Distributions. *Open Journal of Modelling and Simulation*, *7*(1), 79–95. https://doi.org/10.4236/OJMSI.2019.71004

Fagbamigbe, A. F., and Idemudia, E. S. (2016). Survival analysis and prognostic factors of timing of first childbirth among women in Nigeria. *BMC Pregnancy and Childbirth*, *16*(1), 1–12. https://doi.org/10.1186/S12884-016-0895-Y/TABLES/3

Fully Exponential Laplace Approximations for the Joint Modelling of Survival and Longitudinal Data on JSTOR. (n.d.). Retrieved March 16, 2025, from https://www.jstor.org/stable/40247592

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, *1*(4), 465–480. https://doi.org/10.1093/BIOSTATISTICS/1.4.465

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). Joint Models of Longitudinal and Time-to-Event Data with More Than One Event Time Outcome: A Review. *The International Journal of Biostatistics*, *14*(1). https://doi.org/10.1515/IJB-2017-0047

Hsieh, F., Tseng, Y. K., and Wang, J. L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, *62*(4), 1037–1043. https://doi.org/10.1111/J.1541-0420.2006.00570.X

Huang, X., Li, G., Elashoff, R. M., and Pan, J. (2011). A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime Data Analysis*, *17*(1), 80–100. https://doi.org/10.1007/S10985-010-9169-6/METRICS

Köhler, M., Umlauf, N., and Greven, S. (2018). Nonlinear association structures in flexible Bayesian additive joint models. *Statistics in Medicine*, *37*(30), 4771–4788. https://doi.org/10.1002/SIM.7967

Law, N. J. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*, *3*(4), 547–563.

https://doi.org/10.1093/BIOSTATISTICS/3.4.547

Lawson, A. B., Carroll, R., and Castro, M. (2014a). Joint spatial Bayesian modeling for studies combining longitudinal and cross-sectional data. *Statistical Methods in Medical Research*, *23*(6), 611–624. https://doi.org/10.1177/0962280214527383

Lawson, A. B., Carroll, R., and Castro, M. (2014b). Joint spatial Bayesian modeling for studies combining longitudinal and cross-sectional data. *Statistical Methods in Medical Research*, *23*(6), 611–624. https://doi.org/10.1177/0962280214527383

Li, N., Liu, Y., Li, S., Elashoff, R. M., and Li, G. (2021). A flexible joint model for multiple longitudinal biomarkers and a time-to-event outcome: With applications to dynamic prediction using highly correlated biomarkers. *Biometrical Journal*, *63*(8), 1575–1586. https://doi.org/10.1002/BIMJ.202000085

Martins, R., Silva, G. L., and Andreozzi, V. (2016). Bayesian joint modeling of longitudinal and spatial survival AIDS data. *Statistics in Medicine*, *35*(19), 3368–3384. https://doi.org/10.1002/SIM.6937

McHunu, N. N., Mwambi, H. G., Reddy, T., Yende-Zuma, N., and Naidoo, K. (2020). Joint modelling of longitudinal and time-to-event data: An illustration using CD4 count and mortality in a cohort of patients initiated on antiretroviral therapy. *BMC Infectious Diseases*, *20*(1), 1–9. https://doi.org/10.1186/S12879-020-04962-3/TABLES/3

Nguyen, H. T., Vasconcellos, H. D., Keck, K., Reis, J. P., Lewis, C. E., Sidney, S., Lloyd-Jones, D. M., Schreiner, P. J., Guallar, E., Wu, C. O., Lima, J. A. C., and Ambale-Venkatesh, B. (2023). Multivariate longitudinal data for survival analysis of cardiovascular event prediction in young adults: insights from a comparative explainable study. *BMC Medical Research Methodology*, *23*(1), 1–19. https://doi.org/10.1186/S12874-023-01845-4/FIGURES/5

Niekerk, J. van, Bakka, H., and Rue, H. (2021). Competing risks joint models using R-INLA. *Statistical Modelling*, *21*(1–2), 56–71. https://doi.org/10.1177/1471082X20913654

Olayiwola, O. M., Olumuyiwa Ajayi, A., Onifade, O. C., Wale-Orojo, O., and Ajibade, B. (2020). Adaptive cluster sampling with model based approach for estimating total number of hidden COVID-19 carriers in Nigeria. *Statistical Journal of the IAOS*, *36*, 103–109. https://doi.org/10.3233/SJI-200718

Rizopoulos, D. (2012). Joint models for longitudinal and time-to-event data: With applications in R. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*, 1–257. https://doi.org/10.1201/B12208/JOINT-MODELS-LONGITUDINAL-TIME-EVENT-DATA-DIMITRIS-RIZOPOULOS/ACCESSIBILITY-INFORMATION

Rizopoulos, D., Hatfield, L. A., Carlin, B. P., and Takkenberg, J. J. M. (2014a). Combining Dynamic Predictions From Joint Models for Longitudinal and Time-to-Event Data Using Bayesian Model Averaging. *Journal of the American Statistical Association*, *109*(508), 1385–1397. https://doi.org/10.1080/01621459.2014.931236

Rizopoulos, D., Hatfield, L. A., Carlin, B. P., and Takkenberg, J. J. M. (2014b). Combining Dynamic Predictions From Joint Models for Longitudinal and Time-to-Event Data Using Bayesian Model Averaging. *Journal of the American Statistical Association*, *109*(508), 1385–1397. https://doi.org/10.1080/01621459.2014.931236

Rustand, D., van Niekerk, J., Krainski, E. T., Rue, H., and Proust-Lima, C. (2022). Fast and flexible inference for joint models of multivariate longitudinal and survival data using Integrated Nested

Laplace Approximations. *Biostatistics*, *25*(2), 429–448. https://doi.org/10.1093/biostatistics/kxad019

Rustand, D., van Niekerk, J., Rue, H., Tournigand, C., Rondeau, V., and Briollais, L. (2023). Bayesian estimation of two-part joint models for a longitudinal semicontinuous biomarker and a terminal event with INLA: Interests for cancer clinical trial evaluation. *Biometrical Journal*, *65*(4), 2100322. https://doi.org/10.1002/BIMJ.202100322

Sène, M., Bellera, C. A., and Proust-Lima, C. (2014). Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence Titre: Modèles à effets aléatoires partagés pour l'analyse conjointe de données longitudinales et de temps d'événement : application à la prédiction de rechutes de cancer de la prostate. *Journal de La Société Française de Statistique*, *155*(1), 134–155. http://www.sfds.asso.fr/journal

Sun, J., Herazo-Maya, J. D., Molyneaux, P. L., Maher, T. M., Kaminski, N., and Zhao, H. (2019). Regularized Latent Class Model for Joint Analysis of High-Dimensional Longitudinal Biomarkers and a Time-to-Event Outcome. *Biometrics*, *75*(1), 69–77. https://doi.org/10.1111/BIOM.12964

Sweeting, M. J., and Thompson, S. G. (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, *53*(5), 750–763. https://doi.org/10.1002/BIMJ.201100052

Yu, M., Law, N. J., Taylor, J. M. G., and Sandler, H. M. (2004). JOINT LONGITUDINAL-SURVIVAL-CURE MODELS AND THEIR APPLICATION TO PROSTATE CANCER. *Statistica Sinica*, *14*, 835–862.

Yuen, H. P., and Mackinnon, A. (2016). Performance of joint modelling of time-to-event data with time-dependent predictors: An assessment based on transition to psychosis data. *PeerJ*, *2016*(10), e2582. https://doi.org/10.7717/PEERJ.2582/SUPP-4

Zhang, B., Chen, Z., and Albert, P. S. (2012). Latent class models for joint analysis of disease prevalence and high-dimensional semicontinuous biomarker data. *Biostatistics*, *13*(1), 74–88. https://doi.org/10.1093/BIOSTATISTICS/KXR024

Tables

Table1: Estimates of survival model

Coefficient for the variable "sex-female"	-0.711
p-value	<0.001
Wald test	261.1 (p-values <0.001)
Log-rank	302.0 (p-values <0.001)
Concordance statistic	0.757 (standard error = 0.013)

Table 2: Linear mixed-effects model estimates

Variables	**Estimates**	**Pr(>|z|)**
serum bilirubin	0.051	<0.001

serum cholesterol	0.001	<0.001
Albumin	-0.685	<0.001
Alkaline	0.0002	<0.001
SGOT	0.0006	0.031
Platelets	-0.002	0.002
Prothrombin	0.075	0.016
Linear mixed-effects model		
Residual standard deviation		1.91
log-restricted-likelihood		-2842.63
Fixed Effects Coefficients		
Estimated intercept		3.44
Coefficient for drug-D-penicillamine		-0.71
Coefficient for year		0.90
Coefficient for interaction term between drug-D-penicillamine and year		-0.16
Random Effect		
Deviation of the random intercepts across various ID levels		4.02
standard deviation of the random slopes for distinct ID levels		1.11
Correlation coefficient between random slopes and intercepts		0.98

Table 3: Estimates for the proposed and existing Models

	Estimates			**Pr(>|z|)**		
Variables	**Mayor Model**	**Proposed Model 1**	**Proposed Model 2**	**Mayor Model**	**Proposed Model 1**	**Proposed Model 2**
serum bilirubin	0.8707	0.690	0. 773	****	***	***
serum cholesterol			0.018			***
Albumin	-2.533	-1.440	-0.491	***	***	***
Alkaline			0.003			***
SGOT			0.091			**
Platelets			-0.051			***
Prothrombin	2.380	4.237	1.632	**	***	**
log-likelihood			31485.34			

BIC			4226.81			
			64482.31			
R-Square	0.578	0.687	0.891	****	***	***
Adjusted R-Square		0.674	0.885		***	***

* for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$, **** for $p < 0.0001$