Journal of Mathematics and Statistics

Original Research Paper

Joint Modelling of Longitudinal-Time-To-Event with Categorical Variable Indicators of Latent Classes: Application to Tuberculosis Data

Azeez Adeboye*, Mutambayi Ruffin, Ndege James and Qin Yongsong

Department of Statistics, University of Fort Hare, South Africa

Article history
Received: 14-08-2018
Revised: 24-08-2018
Accepted: 08-06-2020

Corresponsing author:
Azeez Adeboye
Department of Statistics,
University of Fort Hare, South Africa
Email: azizadeboye@gmail.com

Abstract: In many clinical and reliability research reports, the outcomes of basic interest is the time to a particular event happens in order to indicate the person's "true" state of health or survival status. Different models have been used to analyze such data separately, but may be unsuitable if the longitudinal and health status measures are correlated. In this study, mixed effect and Cox model of latent class are jointly modelled for the correlation between the covariates, observed and unobserved health status variable with binary latent class indicators. A Bayesian approach for Maximum likelihood estimates is implemented using Markov Chain Monte Carlo (MCMC) techniques. The repeated and survival measures are independently assumed to be a Gaussian process for latent bivariate. The joint model is applied to TB cohort study for the HIV comorbidity effect on event time for Tuberculosis patients. R package is used for curvilinear repeated measures of latent class model and joint latent class models for both repeated measures and survival time event.

Keywords: Bayesian, Binary, Model, Latent, Logistic

Introduction

In various medical researches, the outcomes from the data depend on the time a particular event occurred, for instance time-to-event outcomes in Randomized controlled trials (Argyropoulos and Unruh, 2015). Data in such a way are referred to as time to event (survival) data of a particular interest (Austin *et al*., 2016). However, the event may not be necessarily death, but could be a remission time from a disease, symptoms relief or a disease recurrence. Many of such studies focus on the effect of patients' information on different survival predictions and jointly model the repeated longitudinal with time-to-event measures to develop a prediction models for dynamic event that reestablish over time the evidence of using joint modelling (Andrinopoulou *et al*., 2015).

Time-to-event outcomes can be used to censor a longitudinal data and modelling separately the repeated longitudinal and time-to-event measures, for instance using time-dependent random effect models (Barrett and Su, 2017), linear mixed effects models and Cox regression models (Hickey *et al*., 2016), may sometimes be incompetent to use, which could have a biased effect on the size of the estimates if the two models are correlated (Ibrahim *et al*., 2010). The Cox proportional hazards model is the most commonly utilized for survival model (Cox, 1992) and serves as part of the techniques used for this study. The Cox model has been included with various extensions such as inclusion of random effects (Vaupel *et al*., 1979), longitudinal modeling with outcome-dependent drop-out (Henderson *et al*., 2000), covariates measurement error (Wulfsohn and Tsiatis, 2010), covariates time-dependent (Sweeting and Thompson, 2011), multivariate survival times (Hougaard, 2000) and latent classes for longitudinal measures (Berlin *et al*., 2014).

However, in our study, the Cox model has been included with a classifying latent class indicators with explanatory covariates measured without error. The class membership information is only indirectly available for categorical variables that depends on the latent class regression model distribution, which can be used for multinomial binary indicators of latent classes. This model consists of a multinomial logit model for the covariates with latent class as well as the association of the observed and the latent classes' indicators. Meanwhile, conditionally independence is one of the assumptions for the indicators with given covariates and latent variables, which the probabilities will also to be independent of the covariates.

A joint model assessment for multiple categorical and survival data is established in this study using the latent class logit regression model as a distribution for categorical data and Cox Proportional hazards model as the distribution for survival times data on the conditional latent class. It is of interest to know the extent in which the disease is related to the survival times. One approach to explore this is to examine and establish the measurement model before including the survival time events with binary covariates using a Cox PH model. A semi-parametric PH model is presented to incorporate a variable with latent class as predictor of time-to-event. A binary value is used indirectly to measure the class association and likelihood estimation function is maximized to establish the joint modelling time event data and latent class variables using EM algorithm.

The Methods for the Joint Modelling

The joint modelling comprises of two sub-models: A generalized latent class logit model for binary variables and a Cox proportional hazards model for survival times.

The Generalized Latent Class Logit Model

Y_i = $(Y_{i1},\ldots,Y_{ij})^t$ represents a multiple random variables (random vector) of J binary values for the ith individual in random variables with N individual population, where Y_{ij} is the response of case i on item j of J items and t of a particular class. Let $Z_i = (z_{i1},\ldots,z_{iP})$ denotes (1 X P) row vector of zth covariates for ith individual. Assume that the population comprises of k subpopulations and with latent class variable X_i of the ith individual unobserved. Therefore, the conditional latent class probability distribution for the jth individual is expressed as:

$$\Pr\left(Y_{ij} = y \mid X_i = x\right) = \pi_{xj}\left(1-\pi_{cj}\right)^{1-y}, \tag{1}$$

where, $\pi_x = (\pi_{x1},\ldots, \pi_{xJ})^t$ and $x = (1,\ldots,K)$ are parameters for class-specific probabilities of each of the individuals with value 1. Also, the probability of latent class assumed J individuals are mutually independent event, we have:

$$\Pr\left(Y_{ij} = y \mid X_i = x\right) = \prod_{j=1}^{J} \Pr\left(Y_{ij} = y_{ij} \mid X_i = x\right). \tag{2}$$

However, the generalized logistic model is applied for modelling the association between the covariates (Z_i) and the latent class (X_i) is given as:

$$\Pr\left(X_i = x \mid Z_i\right) = \frac{\exp\left(Z_i\alpha_x\right)}{\sum_{k=1}^{K}\exp\left(Z_i\alpha_k\right)}, \tag{3}$$

where, α_k is a (P×1) scalar vector consisting the parameters for the kth group. Therefore, the distribution of (Y_i,X_i) is expressed as:

$$\begin{aligned}&\Pr\left(Y_i = y_i, X_i = x \mid Z_i\right) = \Pr\left(X_i = x \mid Z_i\right)\\ &\times\Pr\left(Y_i = y_i \mid X_i = x, Z_i\right)\\ &= \frac{\exp\left(Z_i\alpha_k\right)}{\sum_{k=1}^{K}\exp\left(Z_i\alpha_k\right)}\left[\prod_{J=1}^{j}\pi_{xj}^{y_{ij}}\left(1-\pi_{xj}\right)1-y_{ij}\right].\end{aligned}$$

The Cox Proportional Hazards Model

Let $h(t|w_{i1},\ldots,w_{ik})$ signifies the hazard function of ith individual at time t and W_i is the covariates vector comprising of categorical and continuous variables and $\lambda_0(t)$ is signified as baseline hazard function, which corresponds to the intercept term in the regression model. In proportional hazard model, $\lambda_0(t)$ is not specified but positive and assumed a completely non-parametric shape for, which can be expressed as:

$$h\left(t \mid w_i, x_i\right) = \lambda_0(t)\exp\left(w_i\beta + \omega_{xi}\right), \tag{4}$$

where, $W_i = (w_{i1},\ldots,w_{iQ})$ stands for (1×Q) row vector with a corresponding β is the (Q×1) parameter vector. The (K×1) is the parameter vector of ω that contain the effect of the latent class variable X_i on the hazard (ω_1 = 0). As a result of (4), the probability density distribution of the time to events is expressed as:

$$\begin{aligned}&h\left(t \mid w_i, x_i\right) = \lambda_0(t)\exp\left(w_i\beta + \omega_{xi}\right)\\ &\times\exp\left\{-\lambda_0^*(t)\exp\left(w_i\beta + \omega_{xi}\right)\right\},\end{aligned}$$

where, $\lambda_0^*(t) = \int_0^t \lambda_0(s)ds$ is the baseline hazard when integrated. Meanwhile, in a case when the event time is not right–censored but has a non-informative censoring, the probability density distribution (U_i,Δ_i) becomes:

$$\begin{aligned}&h\left(u_i, \delta_i \mid w_i, x_i\right) = \left[\lambda_0\left(u_i\right)\exp\left(w_i\beta + \omega_{xi}\right)\right]^{\delta_i}\\ &\times\exp\left\{-\lambda_0^*\left(u_i\right)\exp\left(w_i\beta + \omega_{xi}\right)\right\}.\end{aligned}$$

However, this study assumes to have non-informative censoring.

The Joint Modelling of Cox PH with Latent Class Mixed Model Indicator

Using a latent class mixed model for regressing (Y_i,X_i) on Z_i and (μ_i,Δ_i) on (z_i,x_i) for Cox Proportional Hazards model, therefore, the joint distribution model of (μ_i,Δ_i) on (z_i,x_i) is expressed as:

$$\Pr\left(\mu_i, \delta_i, y_i, x_i \mid z_i, w_i\right) = \Pr\left(x_i \mid z_i\right)\Pr\left(y_i \mid x_i\right) \times \Pr\left(\mu_i, \delta_i \mid x_i, w_i\right)$$

By integrating the variable with latent class indicators, the distribution for a marginal variables observed (μ_i,Δ_i,Y_i) becomes:

$$\begin{aligned}\Pr\left(\mu_i, \delta_i, y_i, x_i \mid z_i, w_i\right) &= \sum_{x=1}^{k}\Pr\left(x_i \mid z_i\right) \\ &\times \Pr\left(y_i \mid x_i\right)\Pr\left(\mu_i, \delta_i \mid x_i, w_i\right) \\ &= \sum_{x=1}^{k}\left[\frac{\exp\left(z_i\kappa_x\right)}{\sum_{x=1}^{k}\exp\left(z_i\kappa_k\right)}\left\{\prod_{j=1}^{J}\pi_{x_j}^{y_{ij}}\left(1-\pi_{x_j}\right)^{1-y_{ij}}\right\}\right. \\ &\times\left\{\lambda_0\left(\mu_i\right)\exp\left(w_i\beta+\omega_x\right)\right\}^{\delta_i} \\ &\left.\times\exp\left\{-\lambda_0^*\left(\mu_i\right)\exp\left(w_i\beta+\omega_x\right)\right\}\right]\end{aligned}$$

Inference

This study uses Expectation-Maximization (EM) algorithm, which applied to find the maximum likelihood estimates of observed parameters (μ_i,Δ_i,Y_i). This is achieved through iterative method of E-step and M-step. The E-step is used to estimate the expected log-likelihood of the complete data through a conditional estimation on the observed data while M-step is applied in new parameters estimation to maximize the expected log-likelihood.

Suppose θ = (π,κ,$\lambda_0(t)$, β,v) hence, the log-likelihood for the complete data is expressed as:

$$L_{cmpt}\left(\theta;u,\delta,y,x\right) = \sum_{i=1}^{N}L_{i,cmpt}\left(\theta;u_i,\delta_i,y_i,x_i\right),$$

where:

$$\begin{aligned}L_{i,cmpt}\left(\theta;u_i,\delta_i,y_i,x_i\right) &= z_i\kappa_{x_i} - \log\left\{\sum_{x=1}^{k}\exp\left(z_i\kappa_k\right)\right\} \\ &+\sum_{j=1}^{J}\left\{y_{ij}\log\pi_{x_{ij}} + \left(1-y_{ij}\right)\log\left(1-\pi_{x_{ij}}\right)\right\} \\ &+\left[\log\left\{\lambda_0\left(\mu_i\right)\right\} + w_i\beta+\omega_{x_i}\right] - \lambda_0^*\left(\mu_i\right)\exp w_i\beta+\omega_{x_i}\end{aligned}$$

Therefore, the log-likelihood for an observed parameters is expressed as:

$$L\left(\theta;u,\delta,y\right) = \sum_{x\in\{1,\ldots,K\}^N}\exp\left\{L_{i,cmpt}\left(\theta;u,\delta,y,x\right)\right\}$$

From the E-step algorithm, Q_i(θ; θ$^{(\tau)}$) = E(L_{cmpt}(θ;U,Δ,Y,X)|U,Δ,Y;θ$^{(\tau)}$) is estimation of the expected log-likelihood of the complete data with respect to the conditional distribution of (X|U, Δ,Y) to be calculated, where θ$^{(\tau)}$ is the parameter estimate of θ from the rth step.

In the M-step iteration, we maximize the E-step with respect to the parameters obtained in a new setting of the θ′s i.e., Q is maximize as a function of θ. We expressed the posterior distribution of X_i with given parameter data for θ = θ$^{(\tau)}$ to be $\sigma_{i\tau}^{(\tau)}$ = Pr(X_i = x|U, Δ,Y; θ$^{(\tau)}$), c = 1,...,p. Hence, Q and π are maximized as:

$$\pi_{xj}^{(\tau+1)} = \frac{\sum_{i=1}^{N}\sigma_{ix}^{(\tau)}y_{ij}}{\sum_{i=1}^{N}\sigma_{ix}^{(\tau)}}$$

For Q and κ is maximized as:

$$M_{\kappa xn} = \sum_{i=1}^{K}z_{in}\left\{\sigma_{ix}^{(\tau)} - \frac{\exp\left(Z_i\kappa_x\right)}{\sum_{p=1}^{P}\exp\left(Z_i\kappa_P\right)}\right\} = 0$$

Also to maximize (β,v) and Q through ($\beta^{\tau+1}$, $v^{\tau+1}$), we solve:

$$\begin{aligned}M_{\beta_q} &= \sum_{i=1}^{K}z_{iq}\left\{\delta_i - \lambda_0\left(\mu_i\right)\exp\left(w_i\beta\right)\sum_{x=1}^{N}\sigma_{ix}^{(\tau)}\exp\left(v_x\right)\right\} = 0, \\ M_{v_x} &= \sum_{i=1}^{K}\sigma_{ix}^{(\tau)}\left\{\delta_i - \lambda_0\left(\mu_i\right)\exp\left(w_i\beta+v_x\right)\right\} = 0,\end{aligned}$$

where:

$$\lambda_0\left(\mu_i\right) = \sum_{i=1}^{N}\frac{\delta_i I\left(U_i = t\right)}{\sum_{i\in\Re_t}\exp\left(w_i\beta\right)\sum_{x=1}^{N}\sigma_{ix}^{(\tau)}\exp\left(v_x\right)}$$

An update of iterative scheme for (λ_0,β,v) is needed in the M-step, in which the update of λ_0 depends on (β,v). For the update, the following steps are considered:

i. Update of (β,v) is done through Newton-Raphson algorithm
ii. The value of $\lambda_0(t)$ is calculated from $\lambda_0(\mu_i)$ formula
iii. Iteration scheme is carried between (i) and (ii) until convergence

Estimation and Results

Simulation for the estimation of the model was performed to illustrate the method and examine the practicality properties of the proposed joint latent class model and its parameters performances. We generated for

n subjects for longitudinal data from a conditional distribution that is Bernoulli distribution as given by (1) and Weibull distribution for the baseline hazard proportional model to simplify the computation. We described the change in the longitudinal data subject over time using a mixed effects model as a latent class subject-specific variables. We used binary covariates, time continuous covariate and also consider the interaction between them along the intercept term. The covariate z_i represents the ith longitudinal subjects' characteristics.

We denoted the scale and shape parameters of Weibull distribution as Γ. The censoring procedure used is assumed to be uniformly distributed and the given age is transformed linearly to create an effective age distribution function. In each subject, we generated a simulated repeated measurements of n = 200 replicates and k = 2 number of latent class-specific. In each of Monte Carlo simulated data sets, a sample of n = 200 and 500 were generated for a joint modelling defined by $\eta = 0$ with $\mu = 1.0$, $\sigma = 0.80$ and $\lambda_0(\mu) = 1$. The number of repeated nominal time schedule measurements were taken for t_{ij} = 0, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80 and latent class were selected to cater for sufficient and adequate observations in each of the class-specific for easier posterior classification and to show the accuracy of the parameter estimation.

The latent class assumed to be represented as: Logit (π_i) = x_i-0.5. The slope and the intercept were assumed to be distributed uniformly with independent measurement of error (ε_{ij}~N(0,0.1)). The estimates from both longitudinal and survival part were done for the two classes with shape of 1.5 and scale of 20 for class1 and shape 0.5 and scale 10 for class2 respectively. The simulation was performed using R language package.

The Simulation Results

The estimated parameter bias and standard error of the posterior means are reported in Table 1. There was overestimation in the longitudinal measurements of both the intercept and time (t_{ij}) when the sample size is small (n = 200) and the biases estimation for these parameters were more in latent class1 joint model compared to class2. BIC values are used to select the size of the latent classes and used it for the proposed class size K, then compare with the model with smallest BIC. To avoid overestimation and underestimation, we increased the sample size to 500 and found that the two latent class model BIC values are smaller compared to one latent class model and three latent class model respectively. However, this is correctly used to select the number of latent classes with the smallest BIC criterion for simulated datasets. The simulated data of n = 500, is used to evaluate the true value difference in the latent classes joint model. The true value slopes decreases in the same direction for the two latent classes (0.11 against 0.10). Therefore, the important of joint latent class model to identify various sub-groups increases the heterogeneity of the model across the latent class and the estimated standard errors tend to be slightly larger than the true ones, which makes the joint model to be conservative.

Table 1: Simulation results with parameter bias and standard error of the estimates

	Class 1			Class 2		
Parameters	True value	Bias	SE	True value	Bias	SE
N = 200						
Longitudinal part						
Intercept	7.11	0.45	0.34	7.09	0.42	0.28
t_{ij}	0.14	0.31	0.17	0.11	0.30	0.13
Covariate (binary)	0.60	0.74	0.46	0.63	0.71	0.37
t_{ij} *x* covariate (binary)	0.59	-0.40	0.09	0.62	-0.38	0.08
σ_1	1.00	0.50	0.23	1.00	0.46	0.23
σ_2	1.00	0.65	0.30	1.00	0.58	0.31
Survival part						
Covariate (binary)	1.01	0.07	0.42	0.51	0.21	0.09
η	0.54	0.04	0.27	0.52	0.03	0.34
N = 500						
Longitudinal part						
Intercept	7.08	0.41	0.27	7.09	0.40	0.25
t_{ij}	0.21	0.30	0.15	0.10	0.32	0.11
Covariate (binary)	0.60	0.71	0.39	0.62	0.68	0.39
t_{ij} *x* covariate (binary)	0.58	-0.31	0.08	0.60	-0.35	0.05
σ_1	1.00	0.48	0.19	1.00	0.48	0.21
σ_2	1.00	0.59	0.26	1.00	0.55	0.38
Survival part						
Covariate (binary)	1.01	0.06	0.51	1.01	0.08	0.44
η	1.05	0.01	0.16	1.02	0.05	0.32

Estimation Results of Real Data

The joint model is applied to the TB longitudinal retrospective cohort data of all confirmed TB diagnosed in Eastern Cape Province from 2010 to 2015 recorded on monthly and yearly to study the association effect regarding the severe TB prognosis of time event data for elderly patients with those with HIV infection, dialysis, state of immunosuppression and cases of multi-drug resistant TB death risk. A total of 449 patients were included in the study. Due to too small variances of random effects, the age is subtracted from 25 and divided by 30 to decrease the numerical constraints of too large ages in quadratic mixed models. We also considered some covariates such as gender, age, smoking, alcohol use, body weight, smear status, type of TB and diabetes. The time-to-event was measured in days to accommodate the effect of TB prognosis on survival time for TB patients with severity and without.

Model Fit

The Latent Class Mixed Models (LCMM)

The LCMM was implemented to illustrate the quadratic trajectories of TB prognosis factors by assuming that there is correlated random effects for TB patients age-factors functions. The model summary indicates the dataset information, quantity of subjects, number of missing observation deleted, latent classes and parameters. It shows also the number of iterations during the convergence process and criteria to show if the model converged correctly or satisfied. It gives the information about model goodness of fit estimates, which include the maximum log-likelihood value, Akaike Criterion (AIC) and Bayesian Information Criterion (BIC) values. The model also shows the parameter estimated, standard error of estimates, the approximately normal Wald Test statistics and the p-values.

The process of estimating the models with varying latent classes gives the values of log-likelihood estimates, parameters estimates, BIC and posterior proportion of each class. The optimal number of latent classes chosen is two according to the BIC in Table 2. The Bayesian Information criterion for the one (190.18851), two (187.82920, 185.54193, 184.40445) and three-class models (186.89412, 187.32477) respectively, indicating that the two latent class model is better. The proportion of replications from 2c BIC correctly identified the class model 92.1% of the time on the average and that the proportion was close to the probability of choosing the 2 class model randomly out of 1 class and 3 class models. The two-class model corresponds to the most risk TB prognosis factors response patterns, which are diabetes as an indicator and smoking as an indicator. The two groups of the two latent classes as a representative parts for each of the latent classes in terms of TB indicators of Age and gender of the patients (Table 3). The two latent classes may be explained as: (a) a diabetic patients have high prognosis risk factors of TB (b) smoking status of patients is a great risk factors of TB.

The age and gender of patient with TB are treated as predictors of latent class, which are significantly associated with latent class membership with an interaction between the two indicators (Table 3). Interaction of differential item functioning (DIF) is examined in model 2c. The TB factor predictors (age and gender variables) are added separately for each of the five binary factor variables. The differential item functioning estimates and standard errors are summarized in Table 3. The results of the analysis indicate that no significant differential item functioning because the variables from both groups have a different probability results given in each of the responses (the values of z-scores in Table 3 are approximately normally distributed). DIF shows that the items are measuring different abilities for TB indicators subgroups to see if the items are measuring the same way for all TB prognostic factors subgroups. There is interaction of DIF of TB prognostic factors among TB patients from different groups with the same underlying true ability have a different estimation of giving the same results. It showed that Diabetes has highest DIF among the TB prognostic factors for both age and gender (11.33 and 2.18) follow by the alcohol use by the TB patients (0.57 and 1.00). The least of estimation in DIF was found in smoking habit of the TB patients.

The mean subject-specific predictions of the model is presented in Fig. 1a-1d and the goodness of fit statistic of the model support the model using the subject-specific and marginal residual plots in the two-latent class mixed model presented in Figure 4 (Appendix file).

Joint Modelling of Cox PH and Latent Mixed Models

The analysis of joint model is to fit a latent class mixed effects model and survival model using Cox proportional hazard model with baseline hazard functions estimated by Weibull distribution. The models considered having TB prognosis variables (age, gender, diabetes and smoking) as covariates for the latent class mixed model. In the analysis, class-specific quadratic trajectories of TB prognosis risk variables and adjusted for variable TB indicators. We jointly modelled the risk of severe TB prognosis according to diabetes and smoking habit assuming class-specific Weibull baseline risk functions with age and gender of TB patients.

Furthermore, we performed Likelihood Ratio (*LR*) test in comparing the joint model to a longitudinal model with the same covariate effect across latent classes. The analysis of the *LR* test is to show the time effect in longitudinal trajectories, with *p*-value < 0.01, which indicates over the time that the risk of severe TB prognosis has the same pattern validating the use of class specific time effects in the model. We additionally check all covariates (diabetes, smoking status and gender) and found them all to be less than 0.05. In this way, we keep all the covariates that are significant as class-specific in the model and use them to describe the effect of covariates on longitudinal and survival results in class-specific time analysis.

Table 2: Summary table of models estimation process with varying number of latent classes

Models	G	Log-likelihood	No of parameters	BIC	%class1	%class2	%class3
1	1	94.8439	20	190.1885	100.0		
2a	2	103.1036	25	187.8292	7.9	92.1	
2b	2	120.9599	25	185.5419	0.9	99.1	
2c	2	134.0412	25	184.4044	97.1	2.9	
3a	3	141.8356	30	186.8941	3.4	87.3	9.3
3b	3	153.2261	30	187.3247	87.3	3.5	9.2

Table 3: Interaction of differential indicator functioning (DIF) in model (2c)

	TB indicators					
TB prognostic factors	Age25			Gender		
Indicators	Est.	SE	Z-score	Est.	SE	Z-score
Diabetes	11.33	1.952	5.79	2.18	1.079	2.02
Body weight	-0.54	0.285	-1.88	0.56	0.707	0.79
Smoking	-0.19	0.355	-0.55	-0.06	0.441	-0.15
Alcohol use	0.57	0.232	2.45	1.00	0.389	2.56
Smear status	-0.33	0.476	-0.69	-0.50	0.794	-0.63
Type of TB	0.22	0.692	0.31	0.13	0.297	0.44

Table 4: Joint modelling of latent class longitudinal and Cox proportional hazard model

	Parameters	Coef	SE	Wald	P-value
Latent class	Intercept	1.27	0.312	4.07	<0.001
	Age25	0.12	0.006	1.84	<0.001
	Gender	4.12	1.920	2.15	0.032
	Diabetes	6.93	1.833	3.78	<0.001
	Smoking	6.71	1.349	4.94	<0.001
Class 1	Longitudinal part				
	Intercept	1.77	1.026	1.72	0.085
	Age25	0.76	0.747	1.02	0.008
	Gender	0.91	0.683	1.33	0.012
	Age25*gender	0.17	0.238	0.69	0.005
	Survival part				
	Event1 +/-sqrt(Weibull1)	0.04	0.002	2.17	<0.001
	Event1 +/-sqrt(Weibull2)	1.49	0.085	1.72	<0.001
	Event1 SurvPH class1	0.25	0.455	0.55	0.048
	Diabetes	-2.19	0.332	-0.59	0.050
	Smoking	-2.09	0.293	-0.33	0.034
Class 2	Longitudinal part				
	Intercept	0.48	0.438	1.10	0.020
	Age25	0.10	0.318	0.30	0.061
	Gender	0.47	0.371	1.26	0.008
	Age25*gender	0.05	0.318	0.16	0.007
	Survival part				
	Event2 +/-sqrt(Weibull1)	2.67	1.095	2.43	<0.001
	Event2 +/-sqrt(Weibull2)	2.53	0.994	2.54	<0.001
	Event2 SurvPH class2	0.12	0.473	0.44	0.012
	Diabetes	-1.96	1.208	-1.28	0.001
	Smoking	-1.58	0.919	-1.72	0.085

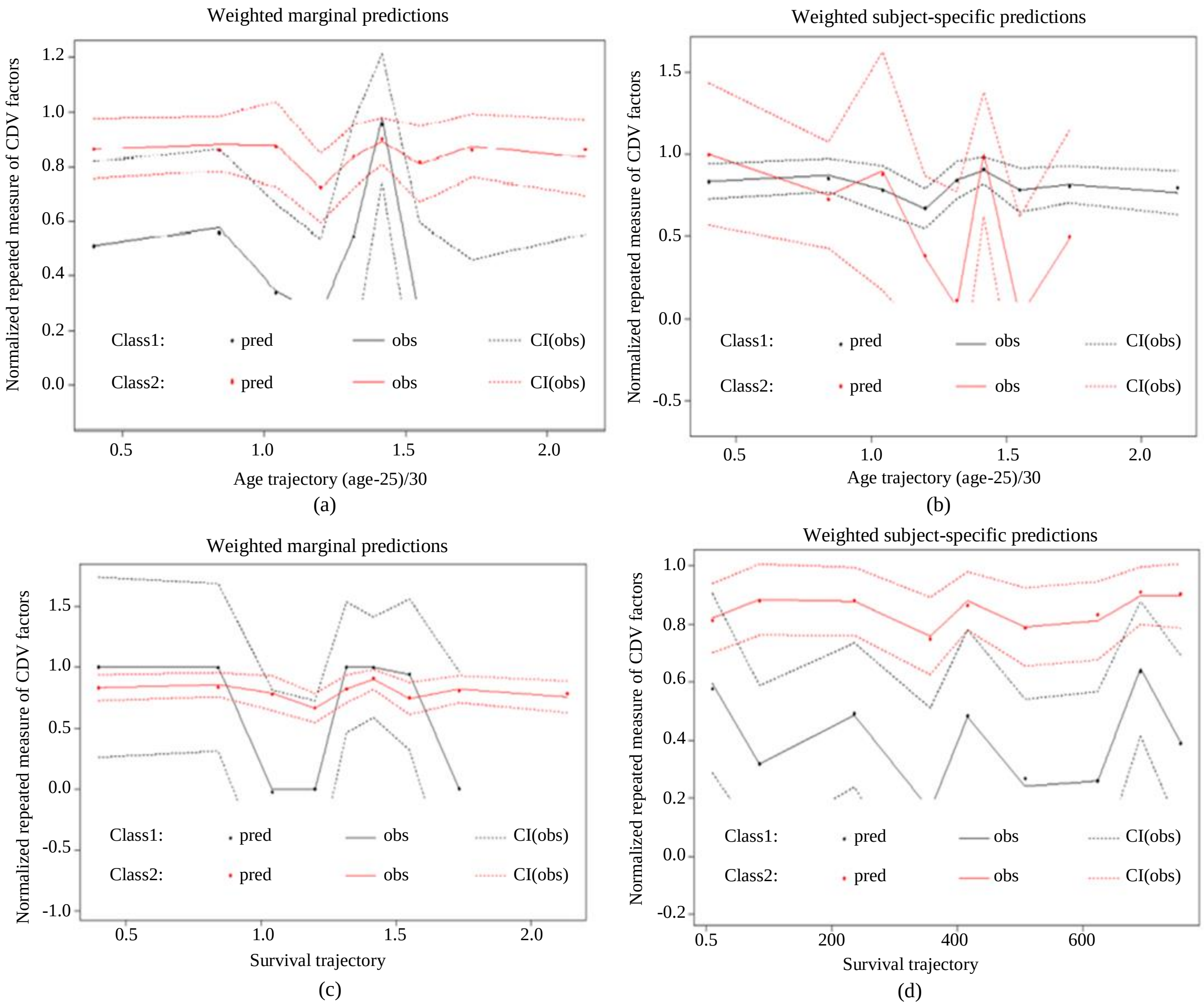


Fig. 1: (a) Weighted mean marginal prediction for two class mixed model; (b) Weighted subject-specific predictions for two-class mixed model; (1) Weighted mean marginal prediction for two class joint model; (d) Weighted subject-specific predictions for two-class joint model

The joint modelling estimates of the two latent class longitudinal and Cox proportional hazard model parameter are shown in Table 4. The estimation results of logistic analysis model indicates that age, gender, diabetes and smoking status are all significant covariates in subjects' classification (P-value <0.05). Male TB patients, who are older in age, diabetic and smoking were less likely to be classified into the class2, signifying that the class2 may comprise of the patients with better health status i.e., may not have the likely course or experience a TB disease. Comparing the estimation in the class2 (0.096) and class1 (0.761) with respect to age (age25) was observed that class1is higher and significantly associated with TB prognosis as the patients getting older but not significant in class2. The value associated with gender among TB patients in class1 is significantly likely to increase the severity of TB disease (0.911) compare to patients in class2. The value of interaction of age and gender of TB patients in class1 was higher compare to class2 and they have a significant effect on event of TB disease prognostic. Two class model possess small biases in parameter estimates and standard error estimates are close to their empirical values.

As for the survival analysis part, diabetes and smoking status are significant to the time-to-event for adults with TB prognosis. This shows that both diabetes and smoking status are significant factors on time-to-event of TB diseases. Patients in the class2 had a better survival rate (HR = 1.128) compare to class1 (HR = 1.284) concerning the effect of severe TB prognosis on the time-to-event for adults with TB disease.

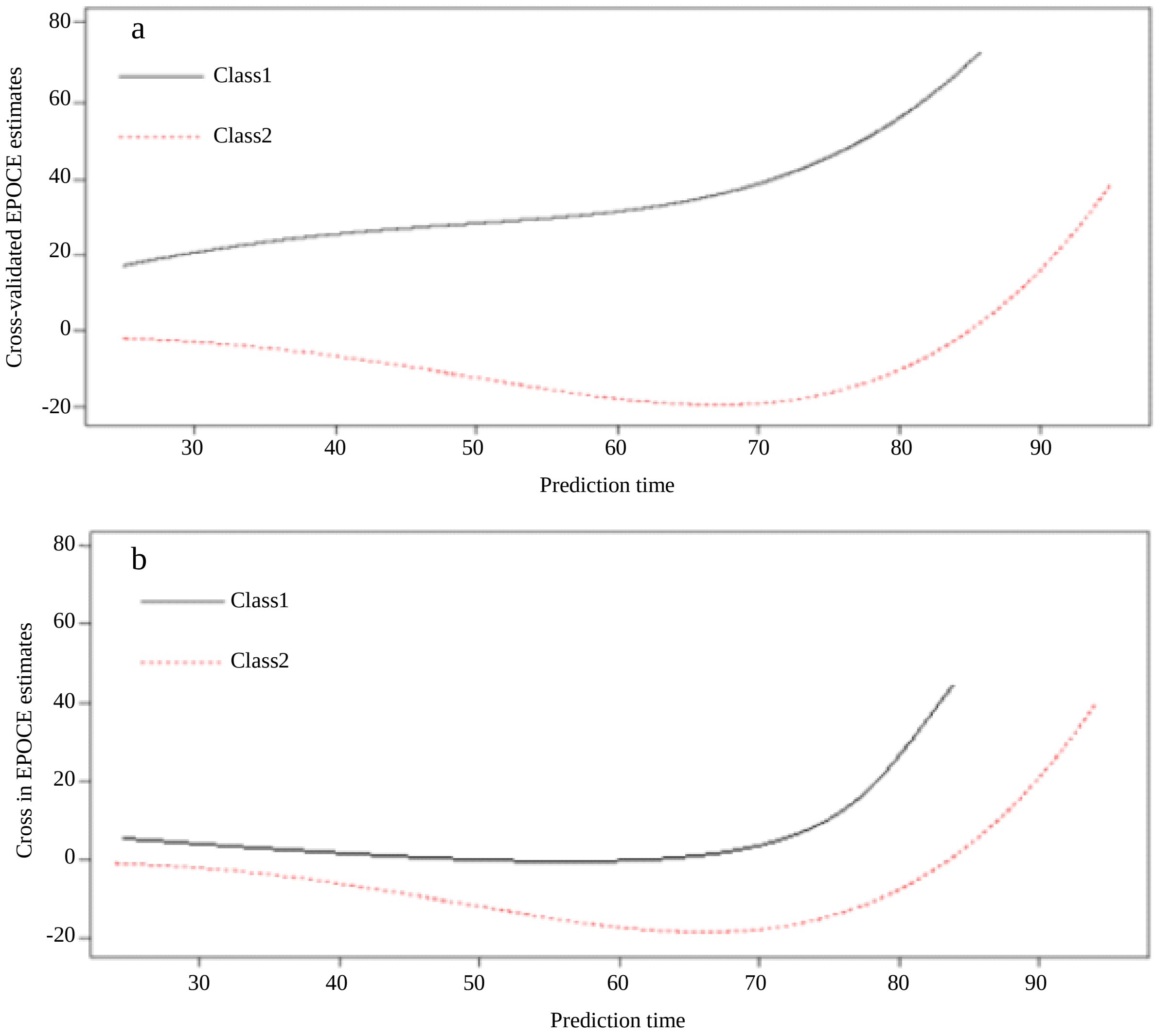


Fig. 2: Above - Predicted Cross-validated EPOCE for joint models with one class. Below -Difference in EPOCE estimates prediction for two latent classes' joint models

Table 5: Posterior classification and probabilities mean table

Posterior probabilities Mean in each class				
	prob1	prob2		
class1	0.8730	0.1270		
class2	0.0309	0.9691		
Posterior classification for longitudinal and time-to-event data				
	class1	class2		
N	41.00	409.00		
%	8.91	91.09		
Variance of the random-effects	Latent class joint model		One class joint model	
	Est.	SE	Est.	SE
Class1	0.09298	0.00413	1.05622	0.08261
Class2	0.10356	0.01289		

The mean of posterior probabilities from the posterior classification (Table 5) is satisfactory. Class1 consist of a posterior 41 subjects (8.91%) while class2 have 409 (91.0%) subjects. The class1 subjects have a mean posterior probability of 87.3% and class2 has a mean posterior probability of 96.9% respectively. Also, the variance of random effects results show that the heterogeneity estimation in one class joint model is larger (1.056) than variance from latent class joint model, suggesting that classification of latent classes reduces the heterogeneity within each class (Table 5).

In summary, part of the objective in this joint modelling is aspect of dynamic prediction of the event for TB patients. We use EPOCE function to compute the predictive ability of our models using the Expected Prognostic Observed Cross Entropy (EPOCE) at different landmark times. The EPOCE is plotted to compare different models to envision the model predictive power at different landmark times. The predictive power difference between the two models is computed using *Diffepoce* function in displaying the associated plot function for both models.

Joint models for TB prognostic factors for two latent classes display a better predictive power showing a lesser EPOCE than a simple one class survival model. Although the two latent class model gives a better goodness-of-fit in terms of BIC with good predictive accuracy, particularly after 80 years of age (Fig. 2).

Discussion

In this study, our model encompasses latent class indicator to jointly model the longitudinal and survival outcomes concurrently. The Cox proportional hazard model has been expanded to encompass variables with latent class measured by binary indicator as factor of time-to-event data. The latent classes stand for different categories of trajectories which were repeatedly measured and assumed to be a normal distribution with given latent class. The proposed model is used to evaluate the likelihood through the mixed effect model and survival processes to discover the underlying trends for the two processes efficiently.

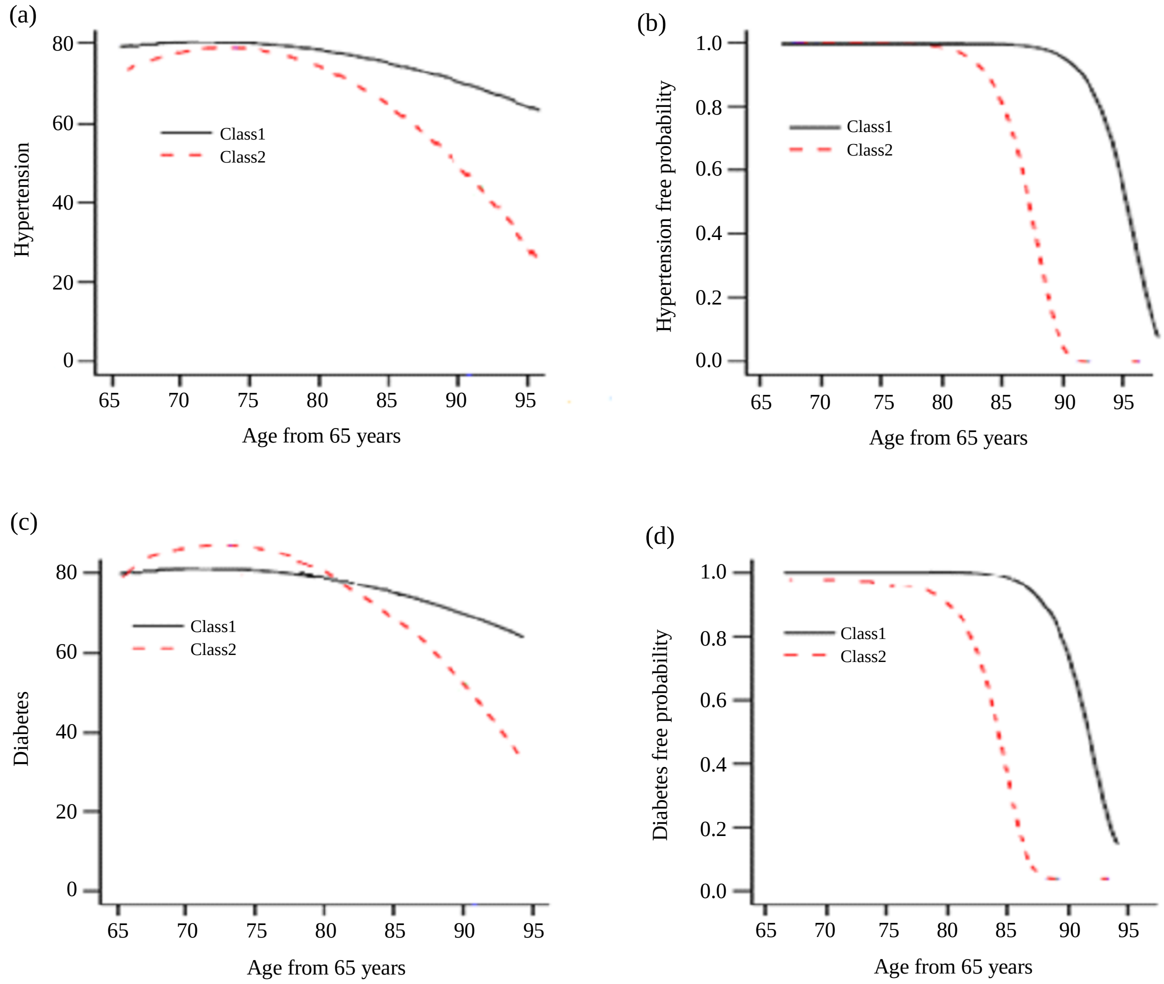


Fig. 3: The predicted trajectories (marginal) of hypertension and diabetes and their associated survival curves in each of the two latent classes for patients with above age of 65 years

The class-specific mean predicted trajectories in each latent class are shown in Fig. 3 according to two prognostic factors (smoking and diabetes). The plots depicts predicted trajectories for both fitted models at a different combination of covariates; male patients who are smoking and above the age of 65years for class1 (dotted lines) and class2 (solid lines). These patients are observed to be at the of severe TB prognostic risk of a prior smoking habit particularly for patients older than 25years. We discovered a relatively increased probability of TB prognosis over the time and the survival of patients having a drug resistant TB is lower in class2 with regard to the patients' age compared to class1.

In finding the estimators for global maximum likelihood, the inference procedures is set up sequentially for initial values in the estimation of the data. In each plots in Fig. 1a and 1b, the dotted lines represent true trajectories and solid lines give the average predicted model trajectories. The below set of dotted and solid lines predicted trajectories are for risk = 0 and above dotted and solid predicted trajectories lines are for risk = 1. The difference between the solid and dashed lines in each pair indicates the model bias. In overall, for one predicted joint model, predicted trajectories are observed with small bias regardless of fitted model (Fig. 1a). For the two interacting predicted joint model, the plots correspond to where the predicted trajectories are provided not only by number of risk, but also by survival status. The plots are largely in bias similar to mixed model plots but smaller in bias regardless of fitted model for the two class joint model. In contrary to the one-predicted joint model case, few bias is also observed for values at the extreme ends when two-interacting predictor joint model is fitted. Summarily, for one or two interacting predictor joint models to be considered, we computed the trajectories prediction from a fitted weighted subject-specific predictions and/or cross-validated EPOCE predictions, regardless of the individual true distribution of estimates differences and little bias expectation. When the sample size is large, trajectories prediction tends to be slightly more efficient when fitting the model.

The model goodness of fit is assessed through the comparison of predicted values against the observed values of the repeated measurement outcomes and plot the martingale residuals of the time-to-event outcomes (Appendix A-Fig. 4). More information about the model fit and subject heterogeneity can be provided through the plots by each classes. However, other forms of estimating baseline hazard functions can be explored such as generalized Gamma distribution and piecewise constant baseline hazard (Liu, 2009; Cox *et al.*, 2007).

Conclusion and Recommendation

The estimation of parameters is jointly analyzed for the latent class mixed effect model with binary indicator and the Cox PH model using EM algorithm. The algorithm is constructed with nonparametric maximum likelihood estimation for the baseline hazard with Weibull distribution for the Cox model. Through simulation, we examined the performance of the joint model with two binary latent class indicators in calculating the joint likelihood and compare it with one class joint model using the likelihood ratio test, variance of the random effects and plotted EPOCE plot to show the comparison of different models to show the predictive power at different landmark times. Additionally, the latent class joint model in our study determined TB prognostic factors, in which the effects can be reduced largely by changing the lifestyle and TB treatment indictors, in which the effects of primary outcomes for TB disease can hardly be improved unless the other risk factors are controlled. The posterior classification from the results illustrated the different trends in the latent classes to show prognostic mean probabilities in each class threshold. The performance of the model from the simulation suggest that the joint estimation with latent class indicators performs better in finite samples. The model is applied on the real dataset to show the advantages of latent class inclusion in the model.

We use EM algorithm for maximum likelihood approach but a Bayesian approach may be of use in the same context. A longitudinal mixed effect data that would allow Poisson indicators can also be looked into as part of the joint model extension. Also, a good idea missing data approach to incorporate a multiple imputation would be suitable for use in a joint modelling approach and useful for future research. A shared random effect to join the longitudinal and survival processes is also a good idea (Liu *et al.*, 2015). It should be noted that sensitivity analysis should also be conducted to estimate the impact of the number of degrees of freedom used for the survival and longitudinal trajectories indicator.

Acknowledgement

The author acknowledge Govan Mbeki Research and Development, University of Fort Hare for partially support the study.

Funding Information

This study was supported by GMRDC under supervisor linked bursary (2017)

Authors contribution

All authors equally contributed in this work.

Ethics

Ethical clearance for the study was obtained from Ethics committee, University of Fort Hare and Department of Health, Eastern Cape chapter, South Africa.

References

Andrinopoulou, E.R., D. Rizopoulos, M.L. Geleijnse, E. Lesaffre and A.J.J.C. Bogers *et al*., 2015. Dynamic prediction of outcome for patients with severe aortic stenosis: Application of joint models for longitudinal and time-to-event data. BMC Cardiovascular Disorders, 15: 28-28. DOI: 10.1186/s12872-015-0035-z

Argyropoulos, C. and M.L. Unruh, 2015. Analysis of time to event outcomes in randomized controlled trials by generalized additive models. PLoS ONE, 10: 1-33. DOI: 10.1371/journal.pone.0123784

Austin, P.C., D.S. Lee and J.P. Fine, 2016. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation, 133: 601-609. DOI: 10.1161/CIRCULATIONAHA.115.017719

Barrett, J. and L. Su, 2017. Dynamic predictions using flexible joint models of longitudinal and time-to-event data. Stat. Med., 36: 1447-1460. DOI: 10.1002/sim.7209

Berlin, K.S., G.R. Parra and N.A. Williams, 2014. An introduction to latent variable mixture modeling (Part 2): Longitudinal latent class growth analysis and growth mixture models. J. Pediatric Psychol., 39: 188-203. DOI: 10.1093/jpepsy/jst085

Cox, C., H. Chu, M.F. Schneider and A. Muñoz, 2007. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statist. Med., 26: 4352-4374. DOI: 10.1002/sim.2836

Cox, D.R., 1992. Regression models and life-tables. J. Royal Stat. Society, 34: 527-541. DOI: 10.1007/978-1-4612-4380-9_37

Henderson, R., P. Diggle and A. Dobson, 2000. Joint modelling of longitudinal measurements and event time data. Biostatistics, 1: 465-480. DOI: 10.1093/biostatistics/1.4.465

Hickey, G.L., P. Philipson, A. Jorgensen and R. Kolamunnage-Dona, 2016. Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. BMC Med. Res. Methodol., 16: 1-15. DOI: 10.1186/s12874-016-0212-5

Hougaard, P., 2000. Analysis of Multivariate Survival Data. 1st Edn., Springer, New York, ISBN-10: 0387988734, pp: 542.

Ibrahim, J.G., H. Chu and L.M. Chen, 2010. Basic concepts and methods for joint models of longitudinal and survival data. J. Clin. Oncol., 28: 2796-2801. DOI: 10.1200/JCO.2009.25.0654

Liu, L., 2009. Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. Statist. Med., 28: 972-986. DOI: 10.1002/sim.3497

Liu, Y., L. Liu and J. Zhou, 2015. Joint latent class model of survival and longitudinal data: An application to CPCRA study. Comput. Stat. Data Anal., 91: 40-50. DOI: 10.1016/j.csda.2015.05.007

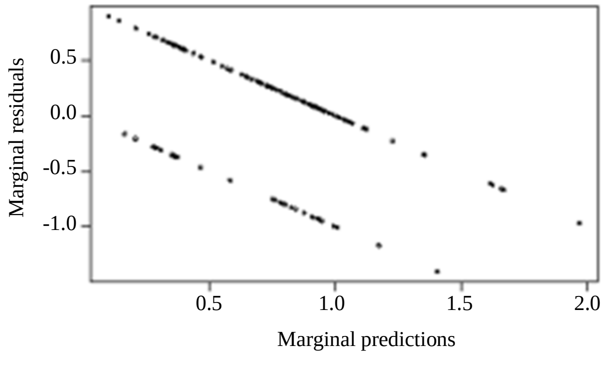
Sweeting, M.J. and S.G. Thompson, 2011. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. Biometrical J., 53: 750-763. DOI: 10.1002/bimj.201100052

Vaupel, J.W., K.G. Manton and E. Stallard, 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography, 16: 439-454. DOI: 10.2307/2061224

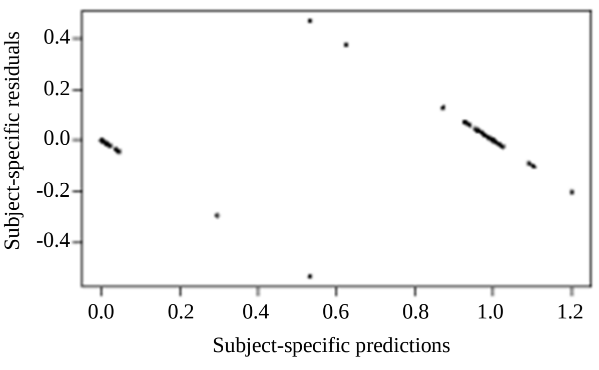
Wulfsohn, M.S. and A.A. Tsiatis, 2010. A joint model for survival and longitudinal data measured with error. Biometrics, 53: 330-339. DOI: 10.2307/2533118

Appendix A

Marginal residuals versus marginal predictions



Subject-specific residuals versus subject-specific predictions



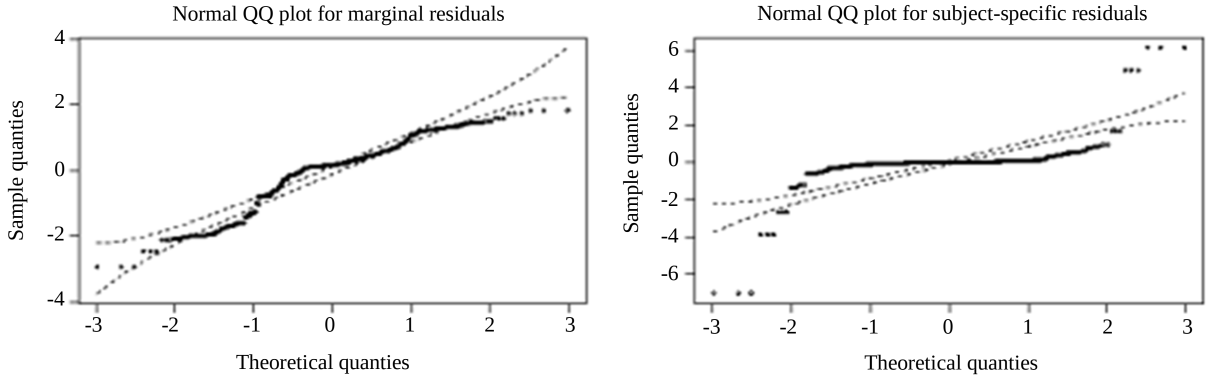


Fig. 4: The model goodness of fit of predicted values against the observed values and plot the martingale residuals of the time-to-event outcomes

R Codes

```
install.packages("lcmm")
library(lcmm)
install.packages("zoo")
install.packages("hlme")
library(multlcmm)
install.packages("multlcmm")
install.packages("jointlcmm")
tb <- read.delim("C:/Users/Azeez/Desktop/tbdata.R")
View(tb)
head(tb)
m1a=hlme(pulse~poly(resrate,degree = 2,raw = TRUE)*sex, random = ~poly(resrate, degree = 2, raw = TRUE), subject = 'id', data = tb, ng = 1)
m1=hlme(pulse~poly(resrate,degree = 2,raw = TRUE)+sex, random = ~poly(resrate, degree = 2, raw = TRUE), subject = 'id', ng = 1, data = tb)
library("NormPsy")
cvd$normpulse = normpulse(cvd$pulse)
tb$Normpulse = Normpulse(tb$pulse)
tb$age=(tb$age - 25)/10
load(tb)
tb$age25 = (tb$age-25)/30
md1=hlme(hpt~age25*pul*sex*history, random=~age25, subject='bmi', ng=1, data=tb)
md2a=hlme(hpt~age25*pul*sex*history, mixture=~age25, random=~age25, classmb = ~pulse+resrate+bloodg+diab, subject = 'bmi', ng=2, data=tb, B=md1)
md2a=hlme(hpt~age25*pul*sex*history, mixture=~age25, random=~age25, classmb = ~arrhy+diab, subject = 'bmi', ng=2, data=tb, B=md1)
md2b=hlme(hpt~age25*pul*sex*history, mixture=~age25, random=~age25, classmb = ~arrhy+diab, subject = 'bmi', ng=2, data=tb, B=random(md1))
md2c=gridsearch (rep = 450, maxiter = 100, minit = md1, hlme(hpt~age25*pul*sex*history, mixture=~age25, random=~age25, classmb = ~arrhy+diab, subject = 'bmi', ng=2, data)
newdata = data.frame(age25=seq(0,5, length = 500), pul=rep(0,500), sex=rep=(0,500), history=rep(0,500), arrhy=rep(0,500), diab=rep(0,500))
postprob(md2c)
plot(md2c)
plot(md2c, which = "fit", var.time = "age25", bty="0" ylab = "hpt", xlab = "(age-25)/30", lwd=1)
plot(md2c, which = "fit", var.time = "age25", ylab = "hpt", xlab = "(age-25)/30", lwd=1, marg = FALSE)
datnew = data.frame(age = seq(25, 95, length = 100))
datnew = data.frame(age = seq(25, 95, length = 449))
datnew$age25 = (datnew$age-25)/30
datnew$diab = 0
diab0 =predictY(md2c, datnew, var.time = "age")
datnew = data.frame(age25 = seq(25, 95, length = 449), pul = seq(0, 1, length=449), sex = seq(0,1, length = 449),
history = seq(0, 1, length=449), arrhy = seq(0, 1, length = 449), diab = seq(0, 1, length = 449))
plot(predicthpt(md2c, datnew, var.time ="age25"), legend.loc = "right", bty = "1")
plot(predict hpt(md2c, datnew, var.time ="age25"), legend.loc = "right", bty = "1")
plot(predict Y(md2c, datnew, var.time ="age25"), legend.loc = "right", bty = "1")
plot(predict (md2c, datnew, var.time ="age25"), legend.loc = "right", bty = "1")
plot ((md2c, datnew, var.time ="age25"), legend.loc = "right", bty = "1")
plot (md2c, datnew, var.time ="age25", legend.loc = "right", bty = "1")
datnew = data.frame(age25 = seq(25, 95, length = 449), pul = seq(0, 1, length=449), sex = seq(0,1, length = 449),
history = seq(0, 1, length=449), arrhy = seq(0, 1, length = 449), diab = seq(0, 1, length = 449))
datnew
dia=predictY(md2c, datnew, var.time="age25")
plot(dia)
plot(dia, lty=1, lwd=2, type="1", col=1.2, ylim=c(25, 95), bty="1", xlab ="age in years", ylab="Diabetic condition",
legend=NULL)
plot(dia, xlab ="age in years", ylab="Diabetic condition", legend=NULL)
plot(dia)
pul1=predictY(md2c, datnew, var.time="age25")
plot(pul1)
sexx=predictY(md2c, datnew, var.time="age25")
sexx
sexy=predictY(md2c, datnew, var.time="hpt")
sexy=predictY(md2c, datnew, var.time="pul")
plot(sexy)
mlin=lcmm(hpt~age25*sex, random=~age25, subject='bmi', data=cvd)
summary(mlin)
joint=Jointlcmm(hpt~age25*pul*sex*history, mixture=~age25, random=~age25, survival = Surv(days,
status)~arrhy+diab, hazard = "Weibull", hazardtype="PH", subject='bmi', ng=2, data=cvd)
summary(joint)
plot(joint, which="fit", var.time="age25", marg = F, break.times = 10, bty ="0", ylab = "hpt", xlab = "age in years")
plot(joint, which="fit", var.time="age25", marg = F, break.times = 10, bty ="2", ylab = "hpt", xlab = "age in years")
plot(joint, which="fit", var.time="age25", marg = F, break.times = 10, ylab = "hpt", xlab = "age in years")
plot(joint)
md1
plot(md1)
plot(md2a)
plot(md1, md2a)
plot(md2b)
plot(md2c)
summarytable(md1, md2a, md2b, md2c)
postprob(md1)
postprob(md2a)
postprob(md2b)
postprob(md2c)
plot(md2c, which="fit", var.time="age25", ylab="hpt", xlab="(age-25)/30", lwd=1, marg=FALSE)
plot(md2b, which="fit", var.time="age25", ylab="hpt", xlab="(age-25)/30", lwd=1, marg=FALSE)
plot(md2a, which="fit", var.time="age25", ylab="hpt", xlab="(age-25)/30", lwd=1, marg=FALSE)
plot(md1, which="fit", var.time="age25", ylab="hpt", xlab="(age-25)/30", lwd=1, marg=FALSE)
plot(md2c, which="fit", var.time="age25", ylab="hpt", xlab="(age-25)/30", lwd=1, marg=FALSE)
dia=predictY(md2c, datnew, var.time="age25")
dia
```

```
plot(dia)
mlin
plot(mlin)
postprob(joint)
plot(joint, which="fit", var.time="days", ylab = "age25", xlab ="survival", lwd=1, marg=FALSE)
jointy=predictY(joint, datnew, var.time ="days")
jointy=predictY(joint, datnew, var.time ="bmi")
jointy=predictY(joint, datnew, var.time ="age25")
jointy=predictY(joint, datnew, var.time ="age25")
summary(jointy)
plot(jointy)
plot(aggregate(cvd$t.stop, by = list(cvd$bmi), FUN=max)[2][ ,1], joint.gap$martingale.res, ylab="", xlab="days", main="Mobility indicators of heart diseases", ylim=c(-5,10))
plot((cvd$t.stop, by = list(cvd$bmi), FUN=max)[2][ ,1], joint.gap$martingale.res, ylab="", xlab="days", main="Mobility indicators of heart diseases", ylim=c(-5,10))
plot(aggregate.data.frame(cvd$t.stop, by = list(cvd$bmi), FUN=max)[2][ ,1], joint.gap$martingale.res, ylab="", xlab="days", main="Mobility indicators of heart diseases", ylim=c(-5,10))
joint1=predictY(joint, datnew, var.time ="hpt")
joint1=predictY(joint, datnew, var.time ="pul")
plot(joint1)
xyplot(log(hpt)~age25|diab, group = status, data = cvd, panel = function(x, y, ...){panel.xyplot(x, y, col = 2, lwd = 2)},)
library(sme)
install.packages("sme")
```