Joint modelling of survival and longitudinal data with informative observation times

Hongsheng Dai*, University of Essex
and
Jianxin Pan†, University of Manchester

Abstract

In this paper, we consider the joint modelling of survival and longitudinal data with informative observation time points. The survival model and the longitudinal model are linked via random effects, for which no distribution assumption is required under our estimation approach. The estimators are shown to be consistent and asymptotically normal. The proposed estimator and its estimated covariance matrix can be easily calculated. Simulation studies and an application to a primary biliary cirrhosis study are also provided.

Keywords: Cox model; informative observation times; log-normal distribution; longitudinal data; multistate models.

1 Introduction

The motivation for this paper arose from a primary biliary cirrhosis (PBC) study (Murtaugh et al., 1994). The PBC is a chronic, fatal, but rare liver disease characterized by inflammatory destruction of the small bile ducts within the liver, which eventually leads to cirrhosis of the liver. Patients often present abnormalities in their blood tests,

*Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK; hdaia@essex.ac.uk

†School of Mathematics, University of Manchester, Oxford Road, Manchester, M13 9PL, UK; Jianxin.Pan@manchester.ac.uk

such as elevated and gradually increased serum bilirubin. The research interest is to study how the drug D-penicillamine (DPCA) affects event times and how the patterns of time courses of bilirubin levels affect death due to PBC. Patients in this study will have their blood tests roughly at 6 months, 1 year, and annually thereafter. Longitudinal measurements (such as bilirubin levels) will be collected at these time points. These predetermined time points are independent of the longitudinal measurements, however, some longitudinal observations may be observed at an 'extra' visit, which is often undertaken unexpectedly because of worsening medical condition. Therefore, such an observation time point is informative to the longitudinal measurement. For survival events, a patient in this study may experience a single event, death/transplant (or censoring); or may experience a death/transplant (or censoring) event and an extra visit to clinic(implying worsening medical condition).

Multiple event models such as multistate models (Andersen and Keiding, 2002; Meira-Machado et al., 2009) are suitable for modelling the extra-visit event and death event. To incorporate the effects of longitudinal measurements, we consider a joint analysis of multiple event models for the survival data and linear mixed effect models for the longitudinal measurements, where the dependency on the informative observation time points is also considered. The sub-models are joint via a common biomarker process. Such joint models for longitudinal data and survival events have been well developed, when the observation times for longitudinal data are non-informative. Henderson et al. (2000) demonstrated the advantage of using a joint modelling approach. Recent developments in this area include Han et al. (2007) for joint models of a longitudinal biomarker and recurrent events; Elashoff et al. (2008) for joint modelling of competing risks models and longitudinal models; and Dantan et al. (2011) for joint analysis of multi-state models and longitudinal models. Note that the longitudinal model in Dantan et al. (2011) involves a change point and they use two different linear models for the longitudinal data before and after the change point. Their model requires that all longitudinal observations are

collected at non-informative time points. In our study, however, the last longitudinal observation may be observed in the extra visit, which is informative. Readers may see Tsiatis and Davidian (2004) for a detailed review of joint modeling of survival events and longitudinal measurements.

The main challenge for the PBC study, which cannot be solved by the existing methods, is the informative observation time points for longitudinal measurements, the 'extra visit'. There has been a vast literature for dealing with informative observation time points for longitudinal data, e.g. Lin et al. (2004); Sun et al. (2007); Liang et al. (2009); Huang et al. (2006); Chen et al. (2015), but these methods focus on longitudinal observations without a terminating survival event or with an independent stopping event. More recent studies focus on longitudinal data with informative observation time points or with informative dropouts. For example Liu et al. (2008) developed a joint random effect model, with the random effect distribution specified, for longitudinal data with informative observation time points and dependent terminal event. Sun et al. (2012) provided a joint analysis of longitudinal data with informative observation times and a dependent terminal event via two latent variables. Their focus is on the effects of the observed covariates rather than how the unobserved biomarker affects the survival events. In their estimation approach, the distribution of the latent variables are unspecified, but the asymptotic covariance matrix is estimated via a Monte Carlo resampling approach. Han et al. (2013) developed a joint modelling approach for longitudinal observations using a semiparametric regression, observation processes and the dropout process using an accelerated failure time model.

To our knowledge, although many methods for joint modelling of multiple survival events and longitudinal data have been developed, joint analysis for multiple survival event and longitudinal data with informative observation times has been hardly studied. In this paper we develop a working-likelihood approach to deal with such problems. This new method has several innovations. First, our method needs neither to impute the latent

random effects nor to integrate out the latent variables from the likelihood. Instead, our method estimates the latent biomarker process via Least Squares estimates, which are actually functions of other unknown parameters in the longitudinal model. Then such Least Squares estimate functions are plugged into the survival model and we can further obtain unbiased estimating equations for all unknown parameters. The proposed method gives a very simple asymptotic covariance matrix estimator, which is easy to compute. Monte Carlo resampling approach is not needed as that in Sun et al. (2012). Second, the new method does not need to specify any distribution for the random effects. This new method is an extension of the corrected score methods in Wang (2006) and Song and Wang (2008). Third, our method can provide not only the effects of the observed covariates on survival events but also how the unobserved biomarker process affects survival event rates.

This paper is organised as follows. The preliminaries and statistical models are introduced in Section 2. Then we introduce the new methodology, and provide the estimating equations, the consistent estimators and the asymptotic normality in Section 3. Simulation studies and an application to the primary biliary cirrhosis study are given in Section 4 and Section 5, respectively. Section 6 gives a brief discussion.

2 Notations and the statistical model

We denote the death event time as T_i, which is usually subject to random censoring by C_i. We can only observe $X_i = \min\{T_i, C_i\}$ and $\delta_i = I[T_i \leq C_i]$. Let $Y_i(t)$ denote the longitudinal process at time t, which is observed intermittently either at time points $t_{i1} < t_{i2} < \cdots < t_{i,n_i}$ (these times are usually planned in advance and are independent of the longitudinal process $Y_i(t)$), except that the last time point may be an extra (random) visit $R_i = t_{i,n_i}$ related to $Y_i(t)$). The observation time R_i means that at this time point the patient visits the clinic unexpectedly. Therefore this 'extra' observation

time point R_i will be informative, for example representing worsening medical condition. We assume that each patient has at most one informative time points, for simplicity and because this is the scenario in the application data set. In general, more than one informative time points could be observed for each patient. The proposed method can be easily extended to such general scenarios (see further discussions in Section 6). For simplicity and without loss of generality, we also assume that there is one time-independent covariate $\boldsymbol{W}_i$, which corresponds to the treatment or other factors. Note that our method is applicable for time-dependent covariates $\boldsymbol{W}_i(t)$, if it can be observed at all time t.

Before presenting the model for $Y_i(t)$, we introduce the counting processes for the death events and the 'extra' visits. We consider a two-state transition model for the death event and the 'extra' visit. For the multistate process, state 0 means alive (the initial state); state 1 means alive but medical condition becomes worse and state 2 means dead. We here only allow transitions from 0 to 1, 0 to 2 and 1 to 2. We define $N^i_{hl}(t) = \#\{$ direct transitions from h to l, in $[0,t]$ for subject $i\}$ and $N_{hl}(t) = n^{-1}\sum_i N^i_{hl}(t)$. A review for multi-state models can be found in Andersen and Keiding (2002).

We consider the longitudinal model

$$Y_i(t) = \mu_i(t) + \alpha\mathcal{H}\{N^i_{01}(t)\} + \epsilon_i(t) \tag{1}$$

where $\mu_i(t)$ is the unobserved biomarker process for subject i before medical condition worsening, $\epsilon_i(t) \sim N(0,\sigma^2)$ and $\mathcal{H}\{N^i_{01}(t)\}$ is a function depending on the counting process related to R_i. This function $\mathcal{H}\{N^i_{01}(t)\}$ is able to model how $Y_i(t)$ changes at or after the informative time point. Without the term $\alpha\mathcal{H}\{N^i_{01}(t)\}$, the random effect $\mu_i(t)$ will be estimated with bias and then this will further distort other parameter estimates. If there are multiple informative times, the function $\mathcal{H}$ can be chosen as the number of informative times within a small neighborhood of t (Sun et al., 2005). In our study, there is at most one informative time. We can choose, for example, $\mathcal{H}\{N^i_{01}(t)\} = I[R_i < t]$.

For simplicity, we denote $\mathcal{H}_i(t) := \mathcal{H}\{N^i_{01}(t)\}$. The random process $\mu_i(t)$ is modelled by $\mu_i(t) = v_{i0} + v_{i1}t + \cdots + v_{iq}t^q$. In our study, we do not specify any distribution assumption on the random effect $\boldsymbol{v}_i = (v_{i0}, v_{i1}, \cdots, v_{iq})$.

Define $\mathcal{F}_i(t)$ as the filtration generated by $\{N^i_{hl}(s), 0 \le s \le t, hl = 01, 02, 12\}$, $\boldsymbol{W}_i$ and $\boldsymbol{v}_i$. The individual rate of the counting processes $N^i_{hl}(s)$ is modelled as a product of a baseline transition rate and a subject specific factor that depends on the covariates and the individual's biomarker process $\mu_i(t) + \alpha\mathcal{H}_i(t_-)$. The biomarker process has a change point at the informative observation time. In summary, we consider the following models,

$$d\Lambda_{hl,i}(t) \quad = \quad d\Lambda_{hl,0}(t) \exp\left\{\boldsymbol{\gamma}'_{hl}\boldsymbol{W}_i + \eta_{hl}(\mu_i(t) + \alpha\mathcal{H}_i(t_-))\right\}. \tag{2}$$

In model (2), parameter $\boldsymbol{\gamma}_{hl}$ shows the effects of time independent covariates such as treatment on the transition rate from state h to state l. Similarly parameter η_{hl} shows the effects of the underlying biomarker process on the transition rate. Notation $d\Lambda_{hl,0}(t)$ means the baseline transition rate from state h to state l.

Equivalently, with the notation $\boldsymbol{\beta}_{hl} = (\boldsymbol{\gamma}_{hl}, \eta_{hl})$ and $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{12}, \boldsymbol{\beta}_{02}, \sigma^2)$, we can also write the above model as

$$\begin{aligned} E[dN^i_{hl}(t)|\mathcal{F}_i(t_-)] \quad &= \quad Q^i_{hl}(t;\boldsymbol{\theta})d\Lambda_{hl,0}(t) \\ Q^i_{hl}(t;\boldsymbol{\theta}) \quad &= \quad \exp\left\{\boldsymbol{\gamma}'_{hl}\boldsymbol{W}_i + \eta_{hl}(\mu_i(t) + \alpha\mathcal{H}_i(t_-))\right\} H^i_{hl}(t). \end{aligned} \tag{3}$$

where $H^i_{01}(t) = H^i_{02}(t) = I[\min\{R_i, X_i\} \ge t]$ and $H^i_{12}(t) = I[X_i \ge t > R_i]$. Note that here if R_i did not occur we let $R_i = \infty$ for convenience. Since we choose $\mathcal{H}\{N^i_{01}(t)\} =$

$I[R_i < t]$, the formula for Q_{hl}^i can also be written as

$$\begin{aligned} Q_{01}^i(t;\boldsymbol{\theta}) &= \exp\left\{\boldsymbol{\gamma}_{01}'\boldsymbol{W}_i + \eta_{01}\mu_i(t)\right\} I[R_i \geq t], \\ Q_{02}^i(t;\boldsymbol{\theta}) &= \exp\left\{\boldsymbol{\gamma}_{02}'\boldsymbol{W}_i + \eta_{02}\mu_i(t)\right\} I[X_i \geq t], \\ Q_{12}^i(t;\boldsymbol{\theta}) &= \exp\left\{\boldsymbol{\gamma}_{12}'\boldsymbol{W}_i + \eta_{12}(\mu_i(t)+\alpha)\right\} I[X_i \geq t > R_i]. \end{aligned} \tag{4}$$

The above joint models are very general comparing to some existing works. If we choose $\mu_i(t) = \mu_0(t)$ to be the same for all subjects and choose $\eta_{hl} = 0$ for all (hl), then the above joint model can be solved using the method in Sun et al. (2005). If we set $\alpha = 0$, $\mu_i(t) = a_i\mu_0(t)$ for some random effect a_i and unknown function $\mu_0(t)$ and we only consider the death (censoring) event, then the above model becomes that in Ding and Wang (2008). The above survival model can also be amended easily to fit the survival data with multiple failure times in Elashoff et al. (2008).

3 The statistical inference

In this section, we focus on the parameter estimations in the models (1) and (3). The challenge here is to deal with the unknown process $\mu_i(t)$. One may specify a particular distribution for $\boldsymbol{v}_i$ and based on this distribution assumption, integrate out the latent variables from the likelihood function (Dantan et al., 2011). However, an inappropriate distribution assumption may distort the final estimation results and if many unknown random effects are involved the numerical integration or EM algorithms will be unstable (Ding and Wang, 2008). Therefore, we here consider an approach without requiring any distribution assumption on $\boldsymbol{v}_i$.

A simple idea is, given all the longitudinal observations for $Y_i(t)$ and α, to replace $\mu_i(t)$ with its Least Squares estimate $\hat{\mu}_i(t;\alpha)$ (as a function of α) in the (partial) likelihood. This simple idea, however, will give biased estimates (Henderson et al., 2000). But

it is possible to use $\hat{\mu}_i(t;\alpha)$ with certain adjustments in the likelihood function or the estimating equations to obtain a consistent estimate. One way of doing this is to use the method based on sufficient statistics in Tsiatis and Davidian (2001) or the corrected score method (Wang, 2006). These methods, however, are only valid when the longitudinal data are collected at noninformative times. This paper uses the log-normal distribution property to *correct* the bias. *Note that for any Gaussian random variable ξ with mean μ_ξ and variance σ_ξ^2 we have that $E[e^\xi] = e^{\mu_\xi+\sigma_\xi^2/2}$. Therefore $e^{\xi-\sigma_\xi^2/2}$ can be used as an unbiased estimate for e^{μ_ξ}.* This log-normal distribution property will help us to find an unbiased estimate for $e^{\mu_i(t)}$, the exponential of the random effect process, which will be part of the proportional hazard model.

3.1 The working likelihood function

We here consider an extension of the corrected score approach. Suppose that subject i has n_i longitudinal observations, $n_i > q+1$. If α is given, we can estimate $\boldsymbol{v}_i$ based on all longitudinal observations of subject i and calculate the predicted value $\hat{\mu}_i(t;\alpha)$. If denoting $\boldsymbol{Y}_i = (Y_i(t_{i1}), \cdots, Y_i(t_{i,n_i}))'$ and $\boldsymbol{\mathcal{H}}_i = (\mathcal{H}_i(t_{i1}), \cdots, \mathcal{H}_i(t_{i,n_i}))'$, we have that $\hat{\mu}_i(t;\alpha) = (1, t, \cdots, t^q)(\mathbf{T}_i'\mathbf{T}_i)^{-1}\mathbf{T}_i'(\boldsymbol{Y}_i - \alpha\boldsymbol{\mathcal{H}}_i)$, where $\mathbf{T}_i$ is the design matrix with n_i rows and $q+1$ columns; the first column has all 1s and the kth column has values t_{ij}^{k-1}, $k \geq 2$. Using standard results from linear regression, given $\mu_i(t)$, the predicted value $\hat{\mu}_i(t;\alpha)$ is normally distributed with mean $\mu_i(t)$ and variance $\sigma_i^2(t) = \sigma^2 b_i(t)$, where

$$b_i(t) = (1, t, \cdots, t^q)(\mathbf{T}_i'\mathbf{T}_i)^{-1}(1, t, \cdots, t^q)'.$$

Recall that $Q_{hl}^i(t;\boldsymbol{\theta})$, defined in (3), is a term used to construct the estimating equations for Cox regression models. Details can be found in Fleming and Harrington (1991) and Andersen et al. (1993). Since Q_{hl}^i is not available (due to the unknown random effects),

we consider using

$$\tilde{Q}^i_{hl}(t;\boldsymbol{\theta}) \;=\; \exp\left\{\boldsymbol{\gamma}'_{hl}\boldsymbol{W}_i + \eta_{hl}(\hat{\mu}_i(t;\alpha) + \alpha\mathcal{H}_i(t)) - \frac{\eta^2_{hl}\sigma^2_i(t)}{2}\right\}\tilde{H}^i_{hl}(t) \qquad (5)$$

instead, where $\tilde{H}^i_{hl}(t) = I[n_i > q+1]H^i_{hl}(t)$. We know that, given $\mu_i(t) = \boldsymbol{v}_i \cdot (1, t, \cdots, t^q)'$, the term $\exp(\eta_{hl}\hat{\mu}_i(t;\alpha))$ follows a log-normal distribution and $E\left[\exp\left(\eta_{hl}\hat{\mu}_i(t;\alpha) - \frac{\eta^2_{hl}\sigma^2_i(t)}{2}\right)|\boldsymbol{v}_i\right] = I[n_i > q+1]\exp(\eta_{hl}\mu_i(t))$. In addition, given $\mathcal{F}_{t_-}$, we also have $E\left[\tilde{Q}^i_{hl}(t;\boldsymbol{\theta})|\mathcal{F}_{t_-}\right] = I[n_i > q+1]Q^i_{hl}(t;\boldsymbol{\theta})$. Therefore it is natural to use $\tilde{Q}^i_{hl}(t;\boldsymbol{\theta})$ to construct the estimating equations. The idea here is to replace $Q^i_{hl}(t;\boldsymbol{\theta})$ by $\tilde{Q}^i_{hl}(t,\boldsymbol{\theta})$ in the standard Cox partial likelihood function and then the estimating equations can be obtained by taking derivatives with respect to parameters α and $\boldsymbol{\beta}_{hl}$ and σ^2.

With the arguments above, we consider the working likelihood function (for the survival sub-model)

$$\tilde{l}(\boldsymbol{\theta}) = \prod_{hl}\prod_{i}\prod_{t}\left[\frac{\exp\left(\boldsymbol{\gamma}'_{hl}\boldsymbol{W}_i + \eta_{hl}(\hat{\mu}_i(t;\alpha) + \alpha\mathcal{H}_i(t)) - \frac{\eta^2_{hl}\sigma^2_i(t)}{2}\right)}{\sum_j \exp\left(\boldsymbol{\gamma}'_{hl}\boldsymbol{W}_j + \eta_{hl}(\hat{\mu}_j(t;\alpha) + \alpha\mathcal{H}_j(t)) - \frac{\eta^2_{hl}\sigma^2_j(t)}{2}\right)\tilde{H}^j_{hl}(t)}\right]^{d\tilde{N}^i_{hl}(t)}.$$

where $d\tilde{N}^i_{hl}(t) = I[n_i > q+1]dN^i_{hl}(t)$. Then the working log-partial likelihood function (for the survival sub-model) is

$$\log\tilde{l}(\boldsymbol{\theta}) = \sum_{hl}\sum_{i}\int\left[\boldsymbol{\gamma}'_{hl}\boldsymbol{W}_i + \eta_{hl}(\hat{\mu}_i(t;\alpha) + \alpha\mathcal{H}_i(t)) - \frac{\eta^2_{hl}\sigma^2_i(t)}{2} - \log\left(\tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})\right)\right]d\tilde{N}^i_{hl}(t). \quad (6)$$

where $\tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta}) := n^{-1}\sum_i \tilde{Q}^i_{hl}(t;\boldsymbol{\theta})$.

On the other hand, since ϵ is normally distributed, we also have the following log-

likelihood function for σ^2 based on the longitudinal sub-model

$$\sum_i I[n_i > q+1] \left[-(n_i - q - 1)\log(\sigma^2) - (\sigma^2)^{-1} \sum_{j=1}^{n_i} (Y_i(t_{ij}) - \hat{\mu}_i(t_{ij};\alpha) - \alpha \mathcal{H}_i(t_{ij}))^2 \right]. \quad (7)$$

Therefore the sum of (6) and (7) will give us the working log-likelihood function.

Note that in Tsiatis and Davidian (2001) σ^2 can be estimated directly based on (7) and then we replace the estimate $\hat{\sigma}^2$ in (6) to estimate the other parameters. This is because in their model $\hat{\mu}_i$ can be estimated directly and (7) does not involve the unknown parameter α. In our study, we need to consider the likelihood as the sum of (6) and (7) and estimate all parameters of $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{12}, \boldsymbol{\beta}_{02}, \sigma^2)$ simultaneously.

3.2 The unbiased estimating equations

We can get the estimating equations, based on the derivative for the log-likelihood. First, we need to introduce the notations $\tilde{\boldsymbol{S}}^{(1)}_{\boldsymbol{\beta};hl}(t,\boldsymbol{\theta})$, $\tilde{S}^{(1)}_{\alpha;hl}(t,\boldsymbol{\theta})$ and $\tilde{S}^{(1)}_{\sigma;hl}(t,\boldsymbol{\theta})$ as the first-order partial derivatives of $\tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}_{hl}$, α and σ^2, respectively. We also introduce the notations $\tilde{S}^{(2)}_{\alpha,\alpha;hl}(t,\boldsymbol{\theta})$, $\tilde{\boldsymbol{S}}^{(2)}_{\alpha,\boldsymbol{\beta};hl}(t,\boldsymbol{\theta})$, $\tilde{S}^{(2)}_{\alpha,\sigma;hl}(t,\boldsymbol{\theta})$, $\tilde{\boldsymbol{S}}^{(2)}_{\boldsymbol{\beta},\boldsymbol{\beta};hl}(t,\boldsymbol{\theta})$, $\tilde{\boldsymbol{S}}^{(2)}_{\boldsymbol{\beta},\sigma;hl}(t,\boldsymbol{\theta})$ and $\tilde{S}^{(2)}_{\sigma,\sigma;hl}(t,\boldsymbol{\theta})$ as the second-order partial derivatives for $\tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}_{hl}$, α and σ^2, respectively. The formulas for these derivatives can be found in Appendix (equations (15), (16) and (17)).

The estimating equations are given by

$$\boldsymbol{U}(\boldsymbol{\theta}) = \left[U_\alpha(\boldsymbol{\theta}), \boldsymbol{U}_{\boldsymbol{\beta}_{01}}(\boldsymbol{\theta})', \boldsymbol{U}_{\boldsymbol{\beta}_{12}}(\boldsymbol{\theta})', \boldsymbol{U}_{\boldsymbol{\beta}_{02}}(\boldsymbol{\theta})', U_\sigma(\boldsymbol{\theta})\right]' = \mathbf{0}$$

for the unknown parameters $\boldsymbol{\theta}$, where

$$
\begin{aligned}
U_\alpha(\boldsymbol{\theta}) &= n^{-1}\sum_{i=1}^{n}\int_0^\infty \sum_{hl}\left[\eta_{hl}(m_i(t)+\mathcal{H}_i(t)) - \frac{\tilde{S}^{(1)}_{\alpha;hl}(t,\boldsymbol{\theta})}{\tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})}\right] d\tilde{N}^i_{hl}(t) - n^{-1}\sum_{i=1}^{n} F_i(\alpha,\sigma^2) \\
\boldsymbol{U}_{\boldsymbol{\beta}_{hl}}(\boldsymbol{\theta}) &= n^{-1}\sum_{i=1}^{n}\int_0^\infty \left[\begin{pmatrix} \boldsymbol{W}_i \\ \hat{\mu}_i(t;\alpha)+\alpha\mathcal{H}_i(t)-\eta_{hl}\sigma_i^2(t)\end{pmatrix} - \frac{\tilde{\boldsymbol{S}}^{(1)}_{\boldsymbol{\beta};hl}(t,\boldsymbol{\theta})}{\tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})}\right] d\tilde{N}^i_{hl}(t) \qquad (8) \\
U_\sigma(\boldsymbol{\theta}) &= n^{-1}\sum_{i=1}^{n}\int_0^\infty \sum_{hl}\left[-\frac{1}{2}\eta_{hl}^2\cdot b_i(t) - \frac{\tilde{S}^{(1)}_{\sigma;hl}(t,\boldsymbol{\theta})}{\tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})}\right] d\tilde{N}^i_{hl}(t) - n^{-1}\sum_{i=1}^{n} E_i(\alpha,\sigma^2)
\end{aligned}
$$

where

$$
\begin{aligned}
E_i(\alpha,\sigma^2) &:= I[n_i > q+1]\left[(n_i-q-1) - \frac{\sum_{j=1}^{n_i}(Y_i(t_{ij})-\hat{\mu}_i(t_{ij};\alpha)-\alpha\mathcal{H}_i(t_{ij}))^2}{\sigma^2}\right]\cdot\sigma^{-2}, \\
F_i(\alpha,\sigma^2) &= I[n_i > q+1]\sum_{j=1}^{n_i} 2(Y_i(t_{ij})-\hat{\mu}_i(t_{ij};\alpha)-\alpha\mathcal{H}_i(t_{ij}))(m_i(t_{ij})+\mathcal{H}_i(t_{ij}))\cdot\sigma^{-2}
\end{aligned}
$$

and $m_i(t)$ is the derivative of $\hat{\mu}_i(t;\alpha)$ with respect to α, given by (14) in Appendix.

Note that given a parameter value $\boldsymbol{\theta}$ we can calculate the function $\boldsymbol{U}(\boldsymbol{\theta})$ given above, since no latent random effects $\boldsymbol{v}_i$ are involved. The expressions of $U_\alpha(\boldsymbol{\theta})$ and $U_\sigma(\boldsymbol{\theta})$ have extra terms $n^{-1}\sum_i F_i$ and $n^{-1}\sum_i E_i$, which are based on the likelihood from the longitudinal data only. We should expect that such estimating equations give consistent estimates, which is shown in the following section.

3.3 Large sample properties of the estimate

Denote the true model parameters as $\boldsymbol{\theta}^*$ and $\Lambda^*_{hl,0}$. Based on the arguments in Section 3.1, i.e. $\tilde{Q}^i_{hl}$ is an unbiased version for Q^i_{hl}, we also have that $\boldsymbol{\theta}^*$ and $\Lambda^*_{hl,0}$ satisfy the model

$$
d\Lambda_{hl,i}(t) = d\Lambda_{hl,0}(t)\exp\left\{\boldsymbol{\gamma}'_{hl}\boldsymbol{W}_i + \eta_{hl}(\hat{\mu}_i(t;\alpha)+\alpha\mathcal{H}_i(t_-)) - \frac{\eta_{hl}^2\sigma_i^2(t)}{2}\right\}. \qquad (9)
$$

Therefore, we only need to show the large sample properties under model (9), which does not involve the unknown random effects.

If we define $\mathcal{G}_i(t)$, as the filtration generated by $\{N^i_{hl}(s), 0 \le s \le t, hl = 01, 02, 12; \boldsymbol{W}_i, \boldsymbol{Y}_i\}$ where $\boldsymbol{Y}_i$ denotes the longitudinal observations, then $\tilde{N}^i_{hl}(t)$ is adapted to $\mathcal{G}_i(t)$ and model (9) can be written as

$$\begin{aligned} E[d\tilde{N}^i_{hl}(t)|\mathcal{G}_i(t_-)] &= d\Lambda_{hl,0}(t)\exp\left\{\boldsymbol{\gamma}'_{hl}\boldsymbol{W}_i + \eta_{hl}(\hat{\mu}_i(t;\alpha) + \alpha\mathcal{H}_i(t_-)) - \frac{\eta^2_{hl}\sigma^2_i(t)}{2}\right\}\tilde{H}^i_{hl}(t) \\ &= d\Lambda_{hl,0}(t)\tilde{Q}^i_{hl}(t;\boldsymbol{\theta})\tilde{H}^i_{hl}(t) \end{aligned} \tag{10}$$

This implies that $dM^i_{hl}(t) = d\tilde{N}^i_{hl}(t) - \tilde{Q}^i_{hl}(t;\boldsymbol{\theta}^*)d\Lambda_{hl,0}(t)$ is a martingale with respect to the filtration $\mathcal{G}_i(t)$. Based on this we can show that the solution of the estimating equations gives a consistent estimate, i.e. $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$, the true parameter value. This is given in Appendix.

To establish the asymptotic normality for the estimator, we first consider the asymptotic normality for the estimating equations. We can show that

$$\sqrt{n}\boldsymbol{U}(\boldsymbol{\theta}^*) \to \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}), \text{ as } n \to \infty$$

for some matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$. This can be easily seen from the facts 1) E_i and F_i $(i = 1, \cdots, n)$ are i.i.d. random variables and 2) the function $\boldsymbol{U}(\boldsymbol{\theta}^*)$ can be rewritten in terms of martingale representation, as

$$\begin{aligned} U_\alpha(\boldsymbol{\theta}^*) &= n^{-1}\sum_{i=1}^n \int_0^\infty \sum_{hl}\left[\eta^*_{hl}(m_i(t) + \mathcal{H}_i(t)) - \frac{\tilde{S}^{(1)}_{\alpha;hl}(t,\boldsymbol{\theta}^*)}{\tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta}^*)}\right]dM^i_{hl}(t) - n^{-1}\sum_{i=1}^n F_i(\alpha,\sigma^2) \\ \boldsymbol{U}_{\boldsymbol{\beta}_{hl}}(\boldsymbol{\theta}^*) &= n^{-1}\sum_{i=1}^n \int_0^\infty \left[\begin{pmatrix} \boldsymbol{W}_i \\ \hat{\mu}_i(t;\alpha) + \alpha\mathcal{H}_i(t) - \eta^*_{hl}\sigma^2_i(t) \end{pmatrix} - \frac{\tilde{\boldsymbol{S}}^{(1)}_{\boldsymbol{\beta};hl}(t,\boldsymbol{\theta}^*)}{\tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta}^*)}\right]dM^i_{hl}(t) \\ U_\sigma(\boldsymbol{\theta}^*) &= n^{-1}\sum_{i=1}^n \int_0^\infty \sum_{hl}\left[-\frac{1}{2}{\eta^*_{hl}}^2 b_i(t) - \frac{\tilde{S}^{(1)}_{\sigma;hl}(t,\boldsymbol{\theta}^*)}{\tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta}^*)}\right]dM^i_{hl}(t) - n^{-1}\sum_{i=1}^n E_i(\alpha^*,{\sigma^*}^2). \end{aligned} \tag{11}$$

Note that the correlation of $dM^i_{hl}(t)$ and E_i (or F_i) is 0, since $E[dM^i_{hl}(t)|\mathcal{G}_i(t_-)] = 0$ and E_i (or F_i) is measurable with respect to $\mathcal{G}_i(t_-)$.

If we define $d\bar{N}_{hl}(t) = n^{-1}\sum_i d\bar{N}^i_{hl}(t)$,

$$\hat{\Lambda}_{hl,0}(t) = \frac{d\bar{N}_{hl}(t)}{\tilde{S}^{(0)}(t,\hat{\boldsymbol{\theta}})}, \tag{12}$$

an estimate for the symmetric matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}$ is given by (the detailed calculation is given in Section 2 of the supplementary file)

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\Sigma}_{\alpha,\alpha} & \hat{\boldsymbol{\Sigma}}_{\alpha,\boldsymbol{\beta}} & \hat{\Sigma}_{\alpha,\sigma} \\ \hat{\Sigma}'_{\alpha,\boldsymbol{\beta}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta},\boldsymbol{\beta}} & \hat{\boldsymbol{\Sigma}}_{\sigma,\boldsymbol{\beta}} \\ \hat{\Sigma}'_{\alpha,\sigma} & \hat{\boldsymbol{\Sigma}}'_{\sigma,\boldsymbol{\beta}} & \hat{\Sigma}_{\sigma,\sigma} \end{pmatrix}$$

where

$$\begin{aligned}
\hat{\Sigma}_{\alpha,\alpha} &= n^{-1}\sum_i \sum_{hl} \int_0^\infty \left[\frac{\tilde{S}^{(2)}_{\alpha,\alpha;hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})} - \frac{\tilde{S}^{(1)}_{\alpha;hl}(t,\hat{\boldsymbol{\theta}})^2}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})^2} \right] d\tilde{N}^i_{hl}(t) + \widehat{\mathrm{Var}}(F_i), \\
\hat{\boldsymbol{\Sigma}}_{\sigma,\sigma} &= n^{-1}\sum_{i=1}^n \sum_{hl} \int_0^\infty \left[\frac{\tilde{S}^{(2)}_{\sigma,\sigma;hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})} - \left(\frac{\tilde{S}^{(1)}_{\sigma;hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})} \right)^2 \right] d\tilde{N}^i_{hl}(t) + \widehat{\mathrm{Var}}(E_i) \\
\hat{\boldsymbol{\Sigma}}_{\alpha,\sigma} &= n^{-1}\sum_{i=1}^n \sum_{hl} \int_0^\infty \left[\frac{\tilde{S}^{(2)}_{\alpha,\sigma;hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})} - \frac{\tilde{S}^{(1)}_{\alpha;hl}(t,\hat{\boldsymbol{\theta}})\tilde{S}^{(1)}_{\sigma;hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})^2} \right] d\tilde{N}^i_{hl}(t) \\
&\quad + \widehat{\mathrm{Cov}}(E_i, F_i)
\end{aligned}$$

with $\widehat{\mathrm{Var}}(F_i)$, $\widehat{\mathrm{Var}}(E_i)$ and $\widehat{\mathrm{Cov}}(E_i, F_i)$ as the sample variances and covariance of $F_i(\hat{\alpha}, \hat{\sigma}^2)$ and $E_i(\hat{\alpha}, \hat{\sigma}^2)$ ($i = 1, \cdots, n$ such that $n_i > q+1$), respectively.

The elements in $\hat{\boldsymbol{\Sigma}}_{\alpha,\boldsymbol{\beta}}$ are given by

$$\hat{\boldsymbol{\Sigma}}_{\alpha,\boldsymbol{\beta}_{hl}} = n^{-1}\int_0^\infty \left[-\begin{pmatrix} \mathbf{0} \\ m_i(t) + \mathcal{H}_i(t) \end{pmatrix} + \left(\frac{\tilde{\boldsymbol{S}}^{(2)}_{\alpha,\boldsymbol{\beta};hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})} - \frac{\tilde{S}^{(1)}_{\alpha;hl}(t,\hat{\boldsymbol{\theta}})\tilde{\boldsymbol{S}}^{(1)}_{\boldsymbol{\beta};hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})^2} \right) \right] d\tilde{N}^i_{hl}(t).$$

The elements in $\hat{\boldsymbol{\Sigma}}_{\sigma,\boldsymbol{\beta}}$ are given by

$$\hat{\boldsymbol{\Sigma}}_{\sigma,\boldsymbol{\beta}_{hl}} = n^{-1}\sum_{i=1}^{n}\int_0^{\infty}\left[\hat{\eta}_{hl}b_i(t) + \left(\frac{\tilde{S}^{(2)}_{\sigma,\boldsymbol{\beta};hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})} - \frac{\tilde{S}^{(1)}_{\sigma;hl}(t,\hat{\boldsymbol{\theta}})\tilde{\boldsymbol{S}}^{(1)}_{\boldsymbol{\beta};hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})^2}\right)\right]d\tilde{N}^i_{hl}(t).$$

The elements in the diagonal blocks of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta},\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_{hl},\boldsymbol{\beta}_{hl}} = n^{-1}\int_0^{\infty}\left[\begin{pmatrix} 0 & 0 \\ 0 & \hat{\sigma}^2 b_i(t)\end{pmatrix} + \frac{\tilde{\boldsymbol{S}}^{(2)}_{\boldsymbol{\beta},\boldsymbol{\beta};hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})} - \left(\frac{\tilde{\boldsymbol{S}}^{(1)}_{\boldsymbol{\beta};hl}(t,\hat{\boldsymbol{\theta}})}{\tilde{S}^{(0)}_{hl}(t,\hat{\boldsymbol{\theta}})}\right)^{\otimes 2}\right]d\tilde{N}^i_{hl}(t)$$

and the other elements in $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta},\boldsymbol{\beta}}$ are 0s.

Then using the first-order Taylor expansion for the estimation equations, we obtain that the asymptotic distribution for $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is $\mathcal{N}(\mathbf{0}, \mathbf{D}^{-1}_{\boldsymbol{\theta}^*}\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}(\mathbf{D}'_{\boldsymbol{\theta}^*})^{-1})$, where $\mathbf{D}_{\boldsymbol{\theta}}$ is the limit of $\partial \boldsymbol{U}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. We then can easily have an estimate for the covariance matrix of $\hat{\boldsymbol{\theta}}$, with $\hat{\boldsymbol{\Sigma}}$ given above and an estimate for $\mathbf{D}_{\boldsymbol{\theta}}$ as $\hat{\mathbf{D}} = \frac{\partial \boldsymbol{U}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, which is given in Section 1 of the supplementary file.

Note that the estimate in (12) for $\Lambda^*_{hl,0}$ is also consistent and $\sqrt{n}(\hat{\Lambda}_{hl,0}(t) - \Lambda^*_{hl,0}(t))$ has asymptotic distribution $N(0, \sigma^2_{hl}(t))$. The variance $\sigma^2_{hl}(t)$ can be estimated as $\hat{\sigma}^2_{hl}(t) = \int_0^t \tilde{S}^{(0}_{hl}(t,\hat{\boldsymbol{\theta}})^{-1}d\hat{\Lambda}_{hl,0}(t) + \boldsymbol{g}_t(\hat{\boldsymbol{\theta}})'\hat{\mathbf{D}}^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\mathbf{D}}')^{-1}\boldsymbol{g}_t(\hat{\boldsymbol{\theta}})$, where $\boldsymbol{g}_t(\boldsymbol{\theta}) = \int_0^t \frac{\partial \tilde{S}^{(0)}(t,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\,\tilde{S}^{(0}_{hl}(t,\boldsymbol{\theta})^{-1}d\hat{\Lambda}_{hl,0}(t)$. This result follows from the standard martingale theory and one may see Chapter 8 of Fleming and Harrington (1991) for more details.

4 Simulation Studies

4.1 Linear random effect process $\mu(t)$

Scenario 1. Simulation studies were carried out to check the performance of the parameter estimators. We choose the longitudinal model with $q = 1$, i.e. $\mu_i(t) = v_{i0} + v_{i1}t$, where random intercept v_{i0} and random slope v_{i1} mimic the subject-specific baseline

disease severity and disease progression rate, respectively (Luo, 2014). The random effects $\boldsymbol{v} = (v_0, v_1)$ were generated from a bivariate normal distribution with mean $(0.3, 0.5)$. The random effects v_0 and v_1 have standard deviation 0.15 and 0.1 respectively and a correlation 0.1. The covariate $\boldsymbol{W}$ is chosen as univariate and generated from a Bernoulli distribution with $p = 0.5$. The baseline hazard rate was chosen as $\lambda_{01}(t) = \exp(-3.0 + 0.2t)$, $\lambda_{02}(t) = \exp(-4.0 + 0.05t)$ and $\lambda_{12}(t) = 1.0$, the censoring variable is $C = 10$ plus an exponential variable with mean 5. This gives a 40% censoring or so. When generating the longitudinal observations we use $\mathcal{H}\{N^i_{01}(t)\} = N^i_{01}(t)$.

The sample sizes were chosen as $n = 200, 400$. All the simulation results in this section are based on 200 Monte Carlo replications. We take longitudinal measurements at every 0.2 unit time when $t <= 1$ and at every 0.5 unit time when $t > 1$. This gives that 95% subjects have more than 2 longitudinal observations ($n_i > 2$) and will contribute to the estimating equations. Even when there are only 85% of subjects having $n_i > 2$, the method still works well. More simulation studies on this and details of how the survival data are generated are provided in the supplementary file.

Table 1 presents the true parameter values, the estimates, Monte Carlo standard deviation (SD), and mean of standard error estimates (SE) and the coverage probabilities for the estimates based on the new methods. We can see from the results that the proposed estimator is practically unbiased. Also the Monte Carlo standard errors agree with the estimated standard errors. This is one of the advantages of the proposed method, which can provide a consistent standard error estimate for the estimated parameters. Existing methods, such as Sun et al. (2012), use bootstrap or other Monte Carlo resampling methods to compute the standard errors which may not be computationally feasible for large data sets. The coverage probabilities of the 95% confidence intervals are also reasonable. As the sample size increases from 200 to 400, the performance of the proposed estimator becomes better.

Table 1 is about here.

The baseline transition rate estimates and the Monte Carlo errors based on the two-hundred simulations are presented in Figure 1. It shows that the baseline estimates are also consistent.

Figure 1 is about here.

Some existing research works use parametric methods to model the random effects and integrate out the unknown random effects from the likelihood function in order to estimate the parameters. For example, Dantan et al. (2011) considered a scenario where a change-point exists for biomarker and used two different linear models (two stages, before and after the change point) to model the longitudinal data. If we rewrite the longitudinal model (1) as

$$\begin{aligned} Y_i(t) &= \mu_i(t) + \epsilon_i(t), \qquad \text{if } t < R_i \\ Y_i(t) &= \mu_i(t) + \epsilon_i(t) + \alpha, \qquad \text{if } t \geq R_i, \end{aligned} \tag{13}$$

then the parametric approach by integrating out the Gaussian random effects can be applied directly. When comparing our method with this approach, we found out that the parametric approach provides similar results to our method, when the random effects are indeed normally distributed (See Table 1). The parametric results even have smaller standard errors, comparing to the working likelihood results. This is not surprising since the parametric method uses the correct Gaussian distribution assumption for the random effects. However, when the random effects are not normally distributed, the parametric methods provide worse results and this is shown in the following simulations studies.

Scenario 2. Now we study the performance of the working likelihood approach under different random effect distributions. We generate v_{i0}, v_{i1} from two independent uniform distributions $U[0.1, 0.5]$ and $U[0.3, 0.7]$. The censoring percentage is about 40%. The simulation results are shown in Table 2. We can see that results based on the working-

likelihood approach are as good as those in Table 1. This is because the proposed working-likelihood method does not require any distribution assumption on the random effect $\boldsymbol{v}$. However, if we use parametric methods with the Gaussian assumption for the random effects, some parameter estimates (those non-zero parameter estimates) have larger bias. This is shown in Table 2.

Table 2 is about here.

Scenario 3. In some cases the parametric approach performs much worse than the proposed working likelihood method. We now consider a more extreme case where v_{i0}, v_{i_1} are generated from mixtures of normal distributions, $v_{i0} \sim 0.5N(0, 0.01)+0.5N(0.1, 0.01)$ and $v_{i1} \sim 0.5N(0.05, 0.01) + 0.5N(0.25, 0.01)$. All other parameter settings remain the same as the previous simulation studies. The censoring percentage is 50% or so. The simulation results are shown in Table 3. We can see that the parametric method provides much larger bias for all parameters. For example, the parametric approach seems not to give a correct estimate for η_{01} and η_{02}. However the working-likelihood approach is very reliable and seems not to be affected by the random effect distribution.

Table 3 is about here.

Scenario 4. For comparison, we also present the result based on $\alpha = 0$ to compare the effects when informative censoring is not taken into account. Table 4 presents the results when the model is misspecified with $\alpha = 0$. We can see the larger bias and poor coverage probability in terms of the estimation for the parameters η_{hl} and σ^2. This is because without α the link process $v_{i0} + v_{i1}t$ will be estimated with bias and therefore its associated parameter η will be estimated with bias.

Table 4 is about here.

4.2 Higher-order polynomials for $\mu(t)$

Theoretically the proposed method require that a large proportion of subjects having $n_i > q+1$, which may limit the applicability of this method. However, practically if $n_i \leq q+1$ we may still estimate the trajectory of $\mu_i(t)$ via a lower-order polynomial. For example, if we choose $q = 3$ then $\mu_i(t)$ should be a 3rd-order polynomial. But for all subjects with $n_i \leq q+1$, we can still estimate $\mu_i(t)$ via a quadratic or linear function. Such an approach only requires that the majority of subjects have no less than 3 repeated observations. Note that if most subjects have only one or two longitudinal observations, it would not make much sense to use a random process to model the biomarker effects. Therefore it is reasonable to focus on problems where most subjects have enough number of longitudinal observations.

In this section we consider a simulation study of $q = 3$. We choose the same true parameter values as before. The random effects $v_{i,0}, v_{i,1}, v_{i,2}, v_{i,3}$ are chosen as independent multivariate Gaussian variables, with means $(-0.2, 0.5, 0.6, -0.04)$ and variances $(0.02, 0.01, 0.002, 0.0001)$. In this simulation study there are only about 50% subjects having more than 4 observations. However, about 95% subjects have more than 2 observations. Therefore, most subjects are included in the working likelihood: some of them (with $n_i = 3$) use linear random effect processes, some of them (with $n_i = 4$) use quadratic random effect processes and some (with $n_i > 4$) use a polynomial function of order 3. We can see from the simulation results in Table 5 that the working-likelihood method still works well.

Table 5 is about here.

5 Data Analysis

Now we apply the proposed approach to the PBC study discussed earlier. In this randomized clinical trial, 158 out of 312 patients took the drug D-penicillamine, whereas the other patients were assigned to a control group. Lab test results such as serum bilirubin were measured at the time of recruitment and at follow-up visits, recorded until death or censoring. The observed event time ranges from 41 to 5225 days. Among the 312 subjects, 125 deaths are observed and the others are censored. The measurement times of serum bilirubin are specified visits at 6 months, 1 year, and annually thereafter. About 85% of patients have no more than 10 longitudinal observations. Following (Luo, 2014), we consider the linear biomarker process $\mu_i(t) = v_{i0} + v_{i1}t$ in case of over fitting. Also about 85% of patients have no less than 3 longitudinal observations, which can contribute to the estimation.

There are 56 patients who have an extra visit. In such an extra visit, patients usually have an abnormal bilirubin values. Several patients' longitudinal observations are plotted in Figure 2. From this plot, we can see that their last bilirubin level is unexpectedly higher than the trend from previous values. These show the worsening medical condition of the patients. Indeed, among these extra-visit patients, 51 of them have observed death.

Figure 2 is about here.

The original data were studied in Fleming and Harrington (1991) based on only baseline covariates, and their conclusion was that the drug D-penicillamine is not effective and some baseline covariates, such as bilirubin, are significant. Ding and Wang (2008) further analysed the data based on the longitudinal observations for bilirubin. They use a joint modelling approach to analyse the survival events and longitudinal data. They also concluded that the drug D-penicillamine is not effective on patient survival but bilirubin levels are significant risk factors.

In our analysis for the survival events, we consider a multi-state model in (3), modeling

transition rate from the initial state (state 0) to the state having the extra visit (state 1), transition rate from having the extra visit to death (state 2) and transition rate from 0 to 2. The covariates in the multi-state model include the longitudinal measurements of serum bilirubin during the follow-up period and the time-independent treatment. When modelling the longitudinal events, we consider model (1). These models will allow us to take into account the effects of the extra visit. Here we mainly focus on explaining the estimated parameter values, shown in Table 6.

Figure 3 is about here.

Table 6 is about here.

The value α is estimated at 0.490 with standard error 0.12, which is significant. This means that at the extra visit, the bilirubin levels are significantly higher than the longitudinal observations obtained before. For four typical patients' longitudinal data, we plot them individually in Figure 3 and show their fitted regression line based on our proposed model. The longitudinal observation at the extra visit is plotted via solid '$\circ$' sign, which is clearly not on the regression line and shows the worsening medical condition. The dotted vertical line in Figure 3 shows the jump of the process at the extra visit and the parameter α can be interpreted as the average of these jumps. Therefore Figure 3 shows that model (1) taking into account the change point would be appropriate.

From Table 6, we have that the estimates $\hat{\gamma}_{02} = -0.067$ (with se 0.26), which is not significant, and $\hat{\eta}_{02} = 0.977$ (with se 0.10), which is significantly unequal to 0. This means that if there is no extra visit, the drug D-penicillamine is not effective on patient survival but bilirubin levels are significant risk factors. This confirms the results in Fleming and Harrington (1991) and Ding and Wang (2008). Based on the new model, however, more results can be achieved. For example, we can also analyse the rate of 'extra visit'. The estimates $\hat{\gamma}_{01} = 0.264$ (with se 0.28), which is not significant, and $\hat{\eta}_{01} = 0.970$ (with se 0.13), which is significantly unequal to 0. This means that the treatment does not affect the rate of the extra visit, but higher bilirubin levels will result

in unexpected visits within a shorter period. This can facilitate clinical managements. On the other hand, $\hat{\gamma}_{12} = -0.543$ (with se 1.42) and $\hat{\eta}_{12} = -0.067$ (with se 0.35), both of which are not significant. This means that if the extra visit happens (medical conditions become worse), neither the treatment nor the bilirubin level will give significant effects on survival. However, the values $\hat{\gamma}_{02} = -0.067$ and $\hat{\gamma}_{12} = -0.543$ might suggest that the treatment has more effects on those patients with worsening conditions. Analysis based on a larger data set is needed to confirm such an argument.

To assess the adequacy of the proposed model, it is straightforward to apply a graphical method based on martingale residuals, similar to Schoenfeld (1982) and Zeng and Cai (2010). The martingale residuals for each subject is given by $\sum_{hl} \int d\tilde{N}^i_{hl}(t) - \tilde{Q}^i_{hl}(t;\boldsymbol{\theta}) d\Lambda_{hl,0}(t)$, which can be calculated easily by using the estimated parameter values. If the multi-state model assumption is reasonable, these residuals should have mean 0 and no correlation with covariates. We find out that the residual mean is 0.157 and residual median is -0.009. In addition the following residual plot shows that there is no relation between residuals and the baseline longitudinal values. Therefore we conclude that our model is appropriate.

Figure 4 is about here.

6 Discussion

In the paper, we have presented a joint model for multi-state event times and longitudinal data with a random process as a link, when there exist informative observation times. Our method does not require any distribution assumption on the random effects. Asymptotically unbiased estimating equations were proposed to obtain parameter estimates and their standard errors. The asymptotic covariance can be easily calculated. Existing corrected score methods or conditional score methods require that the longitudinal process be independent of the data collection times. One contribution of the

new method is that it extends the corrected score method to the cases with longitudinal data collected at informative time points. However, it is not straightforward to extend the conditional score method (Tsiatis and Davidian, 2001), since it is not easy to find a suitable sufficient statistic when the longitudinal process depends on an extra term $\alpha\mathcal{H}_i(t)$. We leave this to a future work.

We here focus on the case where there is only one extra visit. When there are more than one extra visit, the methodology proposed in this paper still works. In such general cases, the counting process $N_{01}(t)$ means the number of extra visits up to time t. Then the three-state transition model proposed in this paper should be revised to a more general multi-state Markov models. We also need to choose different function form $\mathcal{H}(N_{01}(t))$. As Sun et al. (2005) suggested, $\mathcal{H}(N(t))$ can be chosen as the jumps of $N(t)$ at a small neighborhood of t. Under such revised models, the martingale estimation approach in this paper can be applied directly. More research needs to be done to justify the performance of the proposed working likelihood approach in such general scenarios. This is left to future work.

In practice, there may be many subjects having very few or even no longitudinal measurements. Sun et al. (2012) also pointed out this as a challenge and suggested that an inverse probability weight method might work. Such reweighing methods will assign smaller weights to the subjects with few longitudinal observations and larger weights to subjects with more longitudinal observations. From the simulation results provided in the supplementary file (Section 3), we found that our estimators are still very good even if there are about 15% observations have no more than 2 longitudinal observations. But it is worth carrying out further research to study how to incorporate the inverse probability weighted methods into our methods. We leave this as a future work. Nevertheless, the proposed method is preferable to the approach via dealing with unknown random effects using EM algorithm, since the EM algorithm requires a particular distribution assumption for the random effects and will be unstable due to many random effects

included in the model (Ding and Wang, 2008).

In the proposed model, the parameter α is time-independent. If there are more longitudinal observations, this could be generalized to a model with time-dependent parameters. In addition, the random errors $\epsilon_{ij}, j = 1, \cdots, n_i$ could be dependent, following a multivariate normal distribution. It is possible to use the mean-covariance modelling method in Leng et al. (2010) to estimate the covariance matrix for the longitudinal observations. We also leave these as future works.

A Notation for $\boldsymbol{S}^{(1)}$ and $\boldsymbol{S}^{(2)}$ and proof of the consistency

We define

$$m_i(t) := \frac{d\hat{\mu}_i(t;\alpha)}{d\alpha} = -(1, t, \cdots, t^q)(\mathbf{T}_i'\mathbf{T}_i)^{-1}\mathbf{T}_i'\boldsymbol{\mathcal{H}}_i, \quad \frac{d\sigma_i^2(t)}{d(\sigma^2)} = b_i(t). \tag{14}$$

From the definition of $S_{hl}^{(0)}(t, \boldsymbol{\theta})$, we can obtain its derivatives as

$$\begin{aligned}
\tilde{\boldsymbol{S}}_{\boldsymbol{\beta};hl}^{(1)}(t, \boldsymbol{\theta}) &:= \frac{\partial \tilde{S}_{hl}^{(0)}(t, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}_{hl}} = n^{-1} \sum_{i=1}^{n} \begin{pmatrix} \boldsymbol{W}_i \\ \hat{\mu}_i(t;\alpha) + \alpha \mathcal{H}_i(t) - \eta_{hl}\sigma_i^2(t) \end{pmatrix} \tilde{Q}_{hl}^i(t; \boldsymbol{\theta}) \\
\tilde{S}_{\alpha;hl}^{(1)}(t, \boldsymbol{\theta}) &:= \frac{\partial \tilde{S}_{hl}^{(0)}(t, \boldsymbol{\theta})}{\partial \alpha} = n^{-1} \sum_{i=1}^{n} \eta_{hl} \cdot (m_i(t) + \mathcal{H}_i(t)) \cdot \tilde{Q}_{hl}^i(t; \boldsymbol{\theta}) \\
\tilde{S}_{\sigma;hl}^{(1)}(t, \boldsymbol{\theta}) &:= \frac{\partial \tilde{S}_{hl}^{(0)}(t, \boldsymbol{\theta})}{\partial (\sigma^2)} = -n^{-1} \sum_{i=1}^{n} \left(\frac{1}{2} \eta_{hl}^2 b_i(t) \right) \tilde{Q}_{hl}^i(t; \boldsymbol{\theta}).
\end{aligned} \tag{15}$$

We can further work out the second derivatives of $\tilde{S}^{(0)}(t,\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, as

$$\begin{aligned}
\tilde{S}^{(2)}_{\alpha,\alpha;hl}(t,\boldsymbol{\theta}) &:= \frac{\partial^2 \tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})}{\partial\alpha^2} = n^{-1}\sum_{i=1}^{n}\left[\eta_{hl}\cdot(m_i(t)+\mathcal{H}_i(t))\right]^2\cdot\tilde{Q}^i_{hl}(t;\boldsymbol{\theta}) \\
\tilde{\boldsymbol{S}}^{(2)}_{\alpha,\boldsymbol{\beta};hl}(t,\boldsymbol{\theta}) &:= \frac{\partial^2 \tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})}{\partial\alpha\partial\boldsymbol{\beta}_{hl}} = n^{-1}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{0}\\ m_i(t)+\mathcal{H}_i(t)\end{pmatrix}\tilde{Q}^i_{hl}(t;\boldsymbol{\theta}) \qquad (16)\\
&\quad +n^{-1}\sum_{i=1}^{n}\eta_{hl}\cdot(m_i(t)+\mathcal{H}_i(t))\begin{pmatrix}\boldsymbol{W}_i\\ \hat{\mu}_i(t;\alpha)+\alpha\mathcal{H}_i(t)-\eta_{hl}\sigma_i^2(t)\end{pmatrix}\cdot\tilde{Q}^i_{hl}(t;\boldsymbol{\theta}) \\
\tilde{S}^{(2)}_{\alpha,\sigma;hl}(t,\boldsymbol{\theta}) &:= \frac{\partial^2 \tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})}{\partial\alpha\partial(\sigma^2)} = -n^{-1}\sum_{i=1}^{n}\eta_{hl}\cdot(m_i(t)+\mathcal{H}_i(t))\cdot\left(\frac{1}{2}\eta_{hl}^2 b_i(t)\right)\cdot\tilde{Q}^i_{hl}(t;\boldsymbol{\theta}),
\end{aligned}$$

and

$$\begin{aligned}
\tilde{\boldsymbol{S}}^{(2)}_{\boldsymbol{\beta},\boldsymbol{\beta};hl}(t,\boldsymbol{\theta}) &:= \frac{\partial^2 \tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_{hl}^2} \\
&= n^{-1}\sum_{i=1}^{n}\left[\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & -\sigma_i^2(t)\end{pmatrix}+\begin{pmatrix}\boldsymbol{W}_i\\ \hat{\mu}_i(t;\alpha)+\alpha\mathcal{H}_i(t)-\eta_{hl}\sigma_i^2(t)\end{pmatrix}^{\otimes 2}\right]\tilde{Q}^i_{hl}(t;\boldsymbol{\theta}) \\
\tilde{\boldsymbol{S}}^{(2)}_{\boldsymbol{\beta},\sigma;hl}(t,\boldsymbol{\theta}) &:= \frac{\partial \tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})}{\partial\boldsymbol{\beta}_{hl}} = -n^{-1}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{0}\\ \eta_{hl}b_i(t)\end{pmatrix}\tilde{Q}^i_{hl}(t;\boldsymbol{\theta}) \\
&\quad -n^{-1}\sum_{i=1}^{n}\begin{pmatrix}\boldsymbol{W}_i\\ \hat{\mu}_i(t;\alpha)+\alpha\mathcal{H}_i(t)-\eta_{hl}\sigma_i^2(t)\end{pmatrix}\cdot\left(\frac{1}{2}\eta_{hl}^2 b_i(t)\right)\cdot\tilde{Q}^i_{hl}(t;\boldsymbol{\theta}) \\
\tilde{S}^{(2)}_{\sigma,\sigma;hl}(t,\boldsymbol{\theta}) &:= \frac{\partial^2 \tilde{S}^{(0)}_{hl}(t,\boldsymbol{\theta})}{\partial(\sigma^2)^2} = n^{-1}\sum_{i=1}^{n}\left(\frac{1}{2}\eta_{hl}^2 b_i(t)\right)^2\tilde{Q}^i_{hl}(t;\boldsymbol{\theta}). \qquad (17)
\end{aligned}$$

Now we prove the consistency of the estimator $\hat{\boldsymbol{\theta}}$. Denote the log-likelihood, the sum of (6) and (7), as $L_n(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta}) := \lim_n L_n(\boldsymbol{\theta})$ and $\boldsymbol{u}(\boldsymbol{\theta}) = \lim_n \boldsymbol{U}(\boldsymbol{\theta})$. Noticing that $M^i_{hl}(t)$, defined as $dM^i_{hl}(t) = dN^i_{hl}(t) - \tilde{Q}^i_{hl}(t;\boldsymbol{\theta}^*)d\Lambda^*_{hl,0}(t)$ is a martingale with respect to $\mathcal{G}_i(t)$, we have that at the true parameter value $\boldsymbol{\theta}^*$, $\boldsymbol{u}(\boldsymbol{\theta}^*) = \mathbf{0}$. Under certain mild conditions, we will have that $-\partial\boldsymbol{u}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is positive definite and thus $\boldsymbol{\theta}^*$ is the maximum point for the function $L(\boldsymbol{\theta})$.

Therefore the maximum point $\hat{\boldsymbol{\theta}}$ of $L_n(\boldsymbol{\theta})$ converges to the maximum point $\boldsymbol{\theta}^*$ of $L(\boldsymbol{\theta})$. Note that the maximum point $\hat{\boldsymbol{\theta}}$ for $L_n(\boldsymbol{\theta})$ is also the solution of the estimating equations. The consistency is proved.

References

Andersen P. K., Borgan O., Gill R. D. and Keiding N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York Inc..

Andersen P. K. and Keiding N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, **11**: 91-115.

Chen Y., Ning J. and Cai C. (2015) Regression analysis of longitudinal data with irregular and informative observation times. *Biostatistics*, doi: 10.1093/biostatistics/kxv008.

Dantan E., Joly P., Dartigues J-F. and Jacqmin-Gadda H. (2011) Joint model with latent state for longitudinal and multistate data. *Biostatistics*, **12**: 723-736.

Ding J. and Wang J-L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, **64**(2): 546-556.

Elashoff R. M., Li G. and Li N. (2007). An approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in Medicine*, **26**: 2813-2835.

Elashoff R. M., Li G. and Li N. (2008) A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, **64**:762-771.

Flemming T. R. and Harrington D. P. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, Inc..

Murtaugh P. A., Dickson E. R., Van Dam G. M., Malinchoc M., Grambsch P. M.,

Langworthy A. L., Gips C. H. (1994). Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, **20**:126-134.

Han J., Slate E. H. and Pena E. A. (2007). Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Statistics in Medicine*, **26**:5285-5302.

Han M., Song X., Sun L. and Liu L. (2013) Joint modelling of longitudinal data with informative observation times and dropouts. *Statistica Sinica*, Preprint.

Henderson R., Diggle P., Dobson A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**:465-480.

Huang, C. Y., Wang, M. C., and Zhang, Y. (2006) Analyzing panel count data with informative observation times, *Biometrika*, **93**: 763-775.

Leng C., Zhang W. and Pan J. (2010) Semiparametric Mean-Covariance Regression Analysis for Longitudinal Data, *Journal of the American Statistical Association*, **105**: 181-193.

Liang Y., Lu W., and Ying Z. (2009) Joint modeling and analysis of longitudinal data with informative observation times, *Biometrics*, **65**: 377-384.

Lin, H., Scharfstein, D. O., and Rosenheck, D. O. (2004) Analysis of longitudinal data with irregular outcome-dependent follow-up. *Journal of the Royal Statistical Society*, Series B, **66**: 791-813.

Liu, L., Huang X., and O'Quigley J. (2008) Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*, **64**: 950-958.

Luo S. (2014). A Bayesian approach to joint analysis of multivariate longitudinal data and parametric accelerated failure time. *Statistics in medicine*, **33**, 580-594.

Meira-Machado L., de Una-Alvarez J., Cadarso-Suarez C. and Andersen P. K. (2009).

Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, **18**:195-222.

Schoenfeld D. (1982) Partial Residuals for the Proportional Hazards Regression Model. *Biometrika*, **69**:239-241.

Song X. and Wang C.Y. (2008) Semiparametric approaches for joint modeling of longitudinal and survival data with time-varying coefficients. *Biometrics*, **64**:557-566.

Sun J., Park D-H, Sun L. and Zhao X. (2005) Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association*, **100**:882-889.

Sun, J., Sun, L., and Liu, D. (2007) Regression Analysis of longitudinal data in the presence of informative observation and censoring times, *Journal of the American Statistical Association*, **102**: 1397-1406.

Sun, L., Song, X., Zhou J. and Liu, L. (2012) Joint analysis of longitudinal data with informative observation times and a dependent terminal event, *Journal of the American Statistical Association*, **107**: 688-700.

Tisiatis A. A. and Davidian M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, **88**:447-458.

Tsiatis A. A. and Davidian M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**:809-834.

Wang C. Y. (2006). Corrected score estimator for joint modeling of longitudinal and failure time data. *Statistica Sinica*, **16**:235-253.

Zeng D. and Cai J. (2010). A semiparametric additive rate model for recurrent events with an informative terminal event. *Biometrika*, **97**:699-712.

	The new working-likelihood approach							
$n = 200$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
True	**0**.5	**0**.0	**0**.3	**0**.0	**0**.7	-0.5	**0**.3	**0**.09
Estimates	0.505	-0.054	0.314	-0.036	0.671	-0.507	0.279	0.089
SE	0.037	0.184	0.164	0.239	0.227	0.302	0.161	0.0026
SD	0.038	0.190	0.171	0.233	0.239	0.298	0.166	0.0025
CP	0.94	0.94	0.95	0.96	0.93	0.96	0.94	0.94
$n = 400$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
Estimates	0.497	0.028	0.294	0.024	0.707	-0.506	0.289	0.090
SE	0.025	0.128	0.144	0.181	0.203	0.246	0.137	0.0021
SD	0.022	0.122	0.147	0.187	0.196	0.255	0.140	0.0020
CP	0.96	0.96	0.94	0.94	0.95	0.95	0.93	0.94
	The Parametric Gaussian random effect approach							
$n = 400$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
Estimates	0.494	0.021	0.312	0.011	0.685	-0.486	0.290	0.090
SE	0.021	0.098	0.120	0.118	0.199	0.164	0.127	0.0020
SD	0.019	0.102	0.119	0.121	0.195	0.175	0.119	0.0019
CP	0.95	0.96	0.94	0.96	0.94	0.97	0.94	0.95

Table 1: Simulation studies; scenario 1; normal random effects. SE: mean of standard error estimates; SD: Monte Carlo standard deviation of the estimates across the simulated data sets; CP: coverage probability.

	The new working-likelihood approach							
True	**0**.5	**0**.0	**0**.3	**0**.0	**0**.7	-0.5	**0**.3	**0**.09
$n = 200$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
Estimates	0.496	0.002	0.310	0.009	0.680	-0.467	0.274	0.091
SE	0.042	0.138	0.113	0.106	0.150	0.105	0.075	0.004
SD	0.041	0.144	0.117	0.110	0.143	0.107	0.081	0.004
CP	0.93	0.93	0.97	0.93	0.96	0.94	0.95	0.94
$n = 400$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
Estimates	0.505	0.026	0.296	0.005	0.702	-0.504	0.299	0.090
SE	0.031	0.092	0.084	0.081	0.123	0.086	0.061	0.002
SD	0.030	0.094	0.089	0.079	0.126	0.083	0.060	0.002
CP	0.95	0.96	0.96	0.94	0.93	0.96	0.95	0.95
	The Parametric Gaussian random effect approach							
$n = 400$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
Estimates	0.506	-0.010	0.290	0.014	0.728	-0.487	0.286	0.090
SE	0.030	0.082	0.82	0.077	0.119	0.061	0.054	0.002
SD	0.030	0.086	0.085	0.074	0.110	0.067	0.050	0.002
CP	0.94	0.96	0.93	0.97	0.94	0.97	0.94	0.96

Table 2: Simulation studies; scenario 2; uniform random effects. SE: mean of standard error estimates; SD: Monte Carlo standard deviation of the estimates across the simulated data sets; CP: coverage probability.

	The new working-likelihood approach							
True	**0**.5	**0**.0	**0**.3	**0**.0	**0**.7	-0.5	**0**.3	**0**.09
$n = 200$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
Estimates	0.477	0.010	0.312	-0.019	0.675	-0.528	0.276	0.089
SE	0.067	0.228	0.182	0.189	0.192	0.239	0.165	0.009
SD	0.075	0.220	0.191	0.202	0.183	0.244	0.163	0.010
CP	0.96	0.93	0.96	0.97	0.93	0.96	0.94	0.97
$n = 400$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
Estimates	0.490	0.005	0.290	0.020	0.682	-0.516	0.291	0.090
SE	0.052	0.185	0.171	0.187	0.167	0.133	0.149	0.0083
SD	0.057	0.182	0.163	0.179	0.165	0.142	0.147	0.0085
CP	0.92	0.96	0.93	0.95	0.95	0.96	0.95	0.96
	The Parametric Gaussian random effect approach							
$n = 400$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
Estimates	0.534	-0.010	0.232	0.005	0.647	-0.468	0.285	0.080
SE	0.052	0.198	0.113	0.237	0.255	0.171	0.062	0.002
SD	0.054	0.206	0.109	0.242	0.224	0.193	0.071	0.002
CP	0.90	0.96	0.85	0.96	0.90	0.91	0.93	0.96

Table 3: Simulation studies; scenario 2; mixture normal random effects. SE: mean of standard error estimates; SD: Monte Carlo standard deviation of the estimates across the simulated data sets; CP: coverage probability.

$n = 200$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
True	0.5	0.0	0.3	0.0	0.7	-0.5	0.3	0.09
Estimates	-	0.031	0.546	0.015	0.571	-0.527	0.027	0.110
SE	-	0.161	0.207	0.188	0.223	0.235	0.124	0.03
SD	-	0.174	0.192	0.205	0.216	0.217	0.116	0.03
CP	-	0.96	0.60	0.96	0.91	0.94	0.92	0.95

Table 4: Simulation studies; ; scenario 3. SE: mean of standard error estimates; SD: Monte Carlo standard deviation of the estimates across the simulated data sets; CP: coverage probability.

	The new working-likelihood approach							
True	**0**.5	**0**.0	**0**.3	**0**.0	**0**.7	-0.5	**0**.3	**0**.09
$n = 400$	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
Estimates	0.496	0.035	0.296	-0.004	0.695	-0.503	0.293	0.090
SE	0.015	0.112	0.117	0.158	0.196	0.138	0.124	0.0016
SD	0.017	0.110	0.115	0.165	0.191	0.145	0.117	0.0018
CP	0.93	0.95	0.93	0.96	0.95	0.96	0.95	0.95

Table 5: Simulation studies; scenario 5; $q = 3$ and normal random effects. SE: mean of standard error estimates; SD: Monte Carlo standard deviation of the estimates across the simulated data sets; CP: coverage probability.

	α	γ_{01}	η_{01}	γ_{02}	η_{02}	γ_{12}	η_{12}	σ^2
Estimates	0.490	0.264	0.970	-0.067	0.977	-0.543	-0.067	0.108
Std Errors	0.12	0.28	0.13	0.26	0.10	1.42	0.35	0.005

Table 6: Data analysis, estimates and their standard errors.

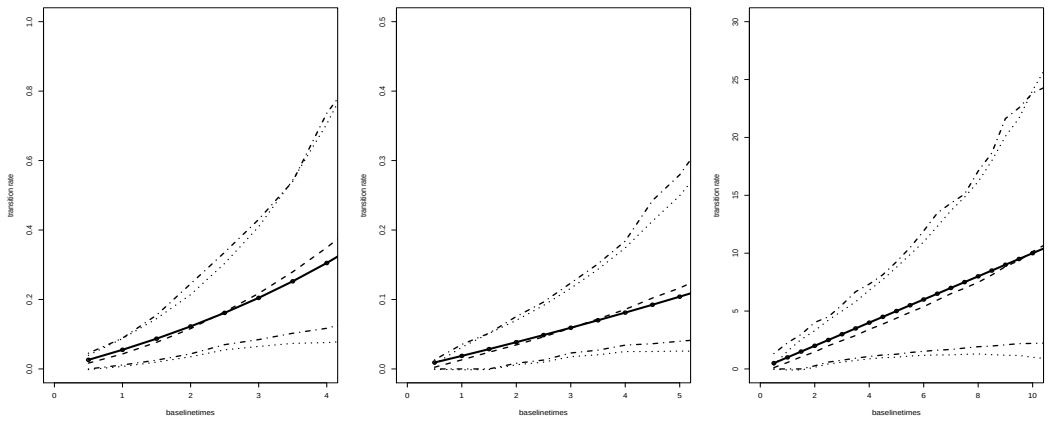

Figure 1: The baseline transition rate estimates for simulation scenario 1. Solid line: the true baseline; dash line: the estimated baseline; dot-dash line: the Monte Carlo error based on the simulation results; dotted line: the mean of the replicated standard error estimates.

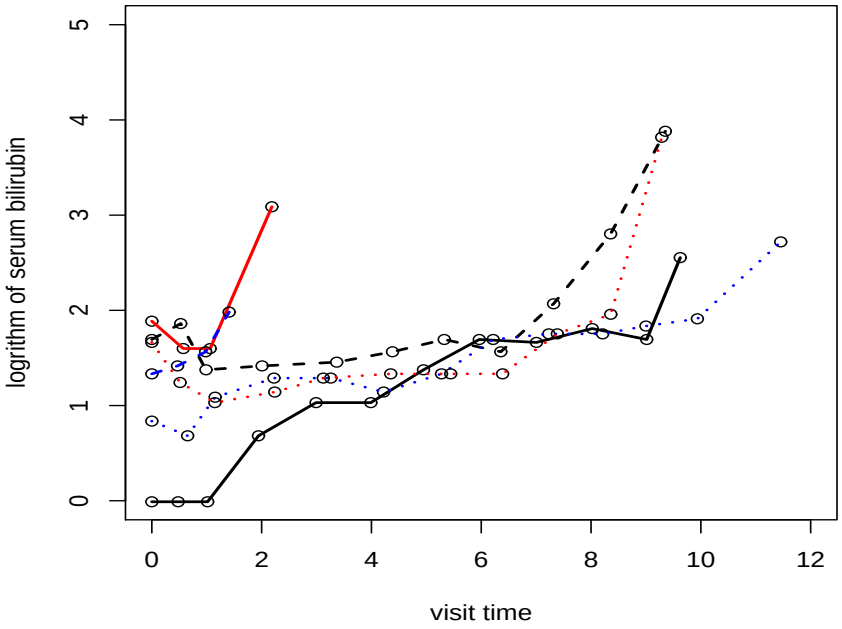

Figure 2: longitudinal observations with the last observation in the extra visit

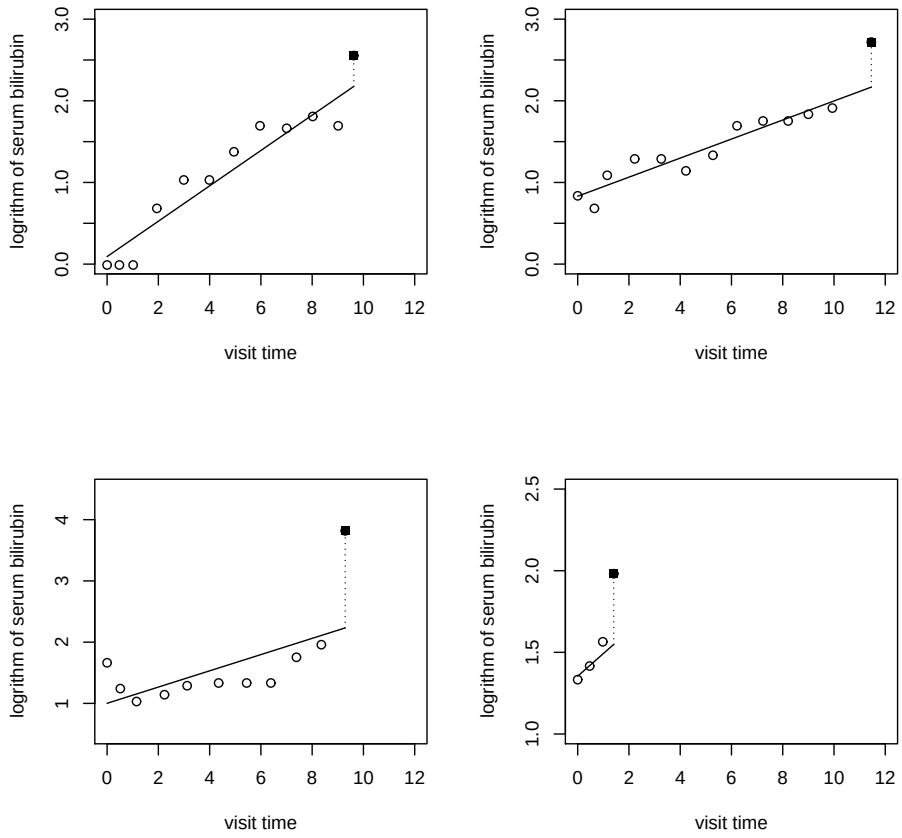

Figure 3: The regression fits for the longitudinal observations, by taking into account that the last visit time is informative.

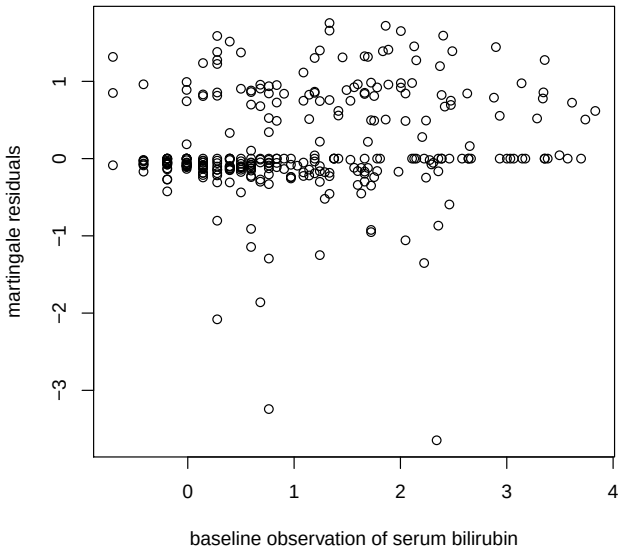

Figure 4: Residual plots for model validation.