Statistics

in Medicine

Research Article

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Model Selection and Diagnostics For Joint Modeling Of Survival And Longitudinal Data With Crossing Hazard Rate Functions

Ka Young Park and Peihua Qiu*

Comparison of two hazard rate functions is important for evaluating treatment effect in studies concerning times to some important events. In practice, it may happen that the two hazard rate functions cross each other at one or more unknown time points, representing temporal changes of the treatment effect. Also, besides survival data, there could be longitudinal data available regarding some time-dependent covariates. When jointly modeling the survival and longitudinal data in such cases, model selection and model diagnostics are especially important to provide reliable statistical analysis of the data, which are lacking in the literature. In this paper, we discuss several criteria for assessing model fit that have been used for model selection, and apply them to the joint modeling of survival and longitudinal data for comparing two crossing hazard rate functions. We also propose hypothesis testing and graphical methods for model diagnostics of the proposed joint modeling approach. Our proposed methods are illustrated by a simulation study and by a real-data example concerning two early breast cancer treatments.

Keywords: crossing hazard rates; joint model; longitudinal data; model diagnostics; model selection; survival data.

Department of Biostatistics, University of Florida, Gainesville, FL 32610

** Correspondence to: Department of Biostatistics, University of Florida, Gainesville, FL 32610*

1. Introduction

Comparison of two hazard rate functions is important in survival data analysis for evaluating treatment effects [1]-[3]. In applications, the two hazard rate functions may cross each other, reflecting temporal changes of treatment effects [4]-[8]. Also, besides survival data, there is often longitudinal data available about some time-dependent covariates [9]-[10]. This paper discusses model selection and model diagnostics when jointly modeling the survival and longitudinal data with crossing hazard rate functions.

The real-data example that motivates this research concerns the two treatments called Cyclophosphamide Epirubicin Fluorouracil (CEF) and Cyclophosphamide Methotrexate Fluorouracil (CMF) of early breast cancer. In a clinical trial study for evaluating the treatment effect of CEF and CMF, 231 patients recruited to the study have their observed survival times between 20-95 months, among which 107 patients received the CEF treatment and the remaining 124 patients received the CMF treatment. The life-table estimates of the hazard rate functions of the two treatment groups are shown in Figure 1. It can be seen that they cross each other around $t = 60$ months. In this example, besides the survival data, longitudinal observations of a quality of life (QOL) index were also recorded for each patient at times when s/he visited the clinic. Therefore, this longitudinal information should also be taken into account when comparing the two crossing hazard rate functions. See Section 4 for a more detailed description of this example.

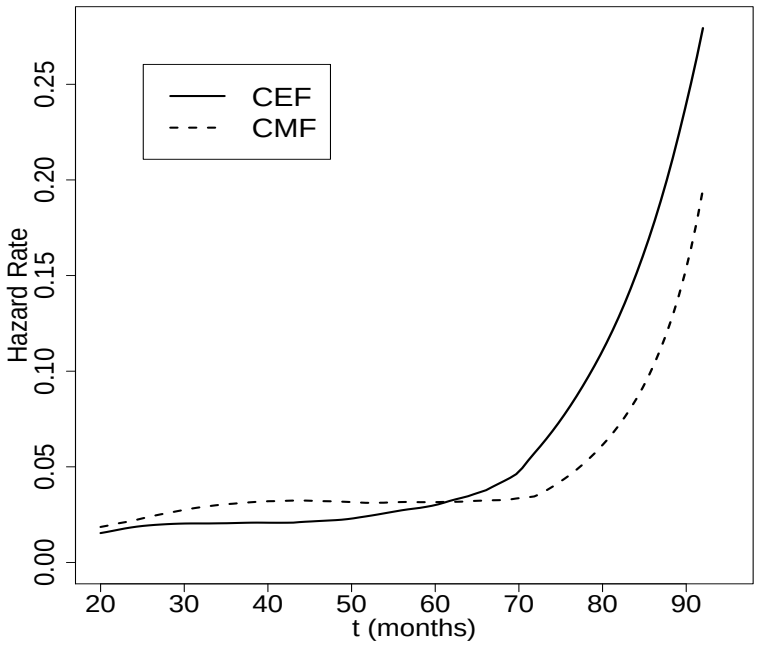

Figure 1. Life-table estimates of the two hazard rate functions of the CEF and CMF treatment groups in the early breast cancer example.

In the literature, there are some existing procedures for comparing two hazard rate functions. Early methods, including

the logrank, Gehan-Wilcoxon, and Peto-Peto tests, among several others (cf., [3], Chapter 7) do not take into account the crossing phenomenon. It has been well demonstrated that these methods are ineffective in comparing two crossing hazard rates, because early differences between the two hazard rates would be canceled out by later differences of opposite signs in their test statistics [11], [7]. To overcome this limitation, several authors, including Fleming et al. [12] and Lin and Wang [6], define their test statistics using absolute or squared differences between the two estimated hazard rates. Some alternative methods handle the crossing hazard rates problem by choosing special weights in the weighted logrank test, which change signs before and after a potential crossing point. See, for instance, [13], [14], and [8] for different weighting schemes. Some other methods employ the modeling approach, by explicitly including the crossing structure of the hazard rates in a model [15], [16], [7], [17]. For recent development on nonparametric estimation of crossing hazard rates, see [18], [19], and the references cited therein. All these existing methods mentioned above for handling the crossing hazard rates problem analyze the survival data only, although some modeling approaches (e.g., [7]) can also accommodate some time-independent baseline information (e.g., a patient's age, gender, etc., at the time when he/she was first included in the study). However, in a medical research (e.g., the early breast cancer example described above), it is common to collect both the time-dependent and time-independent data, besides patients' survival data. In the literature, there are some existing methods for joint modeling the survival and longitudinal data. See, for instance, [9], [10], [20], [21], and the references cited there. But, all these methods do not handle cases when the two hazard rate functions cross each other.

In this paper, we propose a joint modeling procedure to analyze both the survival and longitudinal data in cases when the two hazard rate functions cross each other. Besides model estimation, our focus is on model selection and model diagnostics that are practically important but challenging to discuss in the current problem. To this end, we examine several criteria for assessing fitted models that have been used in the model selection literature, and apply them to the current joint modeling problem. We further discuss hypothesis testing and graphical methods for checking model goodness-of-fit. The rest part of the paper is organized as follows. Section 2 describes our proposed joint model and its estimation procedure, along with our proposed model selection and model diagnostics methods. Section 3 presents a simulation study to investigate their numerical performance. Section 4 demonstrates the proposed methods using the early breast cancer example. Section 5 concludes the article with some concluding remarks.

2. Proposed Method

In this section, we describe our proposed method in three parts. In Subsection 2.1, the proposed joint model and its estimation are first discussed. Then, some model selection criteria are described in Subsection 2.2. Finally, our proposed methods for model diagnostics are discussed in Subsection 2.3.

2.1. Joint modeling procedure and its estimation

Assume that there are n subjects involved in a study. For the survival data of the ith subject with $i = 1, 2, \ldots, n$, let $O_i = \min(T_i, C_i)$ denote the observed survival time and $\Delta_i = I\{T_i \leq C_i\}$ denote the censoring indicator, where T_i is the true survival time and C_i is the censoring time of the ith subject. For the longitudinal data of the ith subject, assume that the response variable $Y(t)$ is observed at n_i time points and it follows the linear mixed effects model

$$\begin{aligned} Y(t_{ij}) &= \boldsymbol{X}(t_{ij})'\boldsymbol{\beta} + \boldsymbol{Z}(t_{ij})'\boldsymbol{b}_i + \varepsilon_{ij} \\ &=: M(t_{ij}) + \varepsilon_{ij}, \qquad \text{for } j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, n, \end{aligned} \tag{1}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the covariates of the fixed effects and the random effects, respectively, $\boldsymbol{\beta}$ is the vector of the fixed effects coefficients, $\boldsymbol{b}_i \overset{iid}{\sim} N(\boldsymbol{0}, \Sigma_b)$ is the vector of the random effects coefficients, and $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ are the random errors. In model (1), it is routinely assumed that the random errors are independent of the random effects. For the survival data, we assume that the survival model is

$$\lambda\left(t|M(t), g\right) = \lambda_0(t) \exp\left\{\psi M(t) + \phi\left(t - \gamma\right) g + \boldsymbol{W}'\boldsymbol{\eta}\right\}, \tag{2}$$

where $\lambda_0(t)$ is the baseline hazard rate, $M(t)$ is the mean component of the longitudinal response $Y(t)$ defined in model (1), g is the group indicator that equals 1 if a subject is in the first treatment group and 0 otherwise, ψ, ϕ and γ are unknown coefficients, $\boldsymbol{W}$ is the vector of some extra covariates, and $\boldsymbol{\eta}$ is a vector of coefficients. By model (2), the log-ratio of the hazard rate functions of the two treatment groups is

$$\log\left[\frac{\lambda\left(t|M(t), g = 1\right)}{\lambda\left(t|M(t), g = 0\right)}\right] = \phi\left(t - \gamma\right),$$

which changes signs at $t = \gamma$. Therefore, by using model (2), we actually assume that the two hazard rate functions cross at γ given $M(t)$. It is worth mentioning that if the value of $M(t)$ is not given and $M(t)$ depends on the group indicator g, then the crossing point of the two hazard rate functions is usually not γ. For ease of understanding, let us re-write $M(t)$ as $M(t, g)$. Then, the log-ratio of the hazard rate functions in such cases is $\psi[M(t, 1) - M(t, 0)] + \phi(t - \gamma)$, and the crossing point should be the root of the equation $\psi[M(t, 1) - M(t, 0)] + \phi(t - \gamma) = 0$. See expression (11) and the related discussion in Section 4 for an example.

For models (1) and (2), we would like to make several remarks. First, model (2) assumes a linear pattern of the log hazard ratio around the crossing point γ. This can be generalized in several different ways. For instance, the term $\phi\,(t-\gamma)\,g$ in model (2) can be generalized to $\phi\,[\mathrm{BC}_\alpha(t)-\mathrm{BC}_\alpha(\gamma)]\,g$, where $\mathrm{BC}_\alpha(t)$ is the Box-Cox transformation of t and α is a parameter. More specifically, $\mathrm{BC}_\alpha(t)=(t^\alpha-1)/\alpha$ if $\alpha\neq 0$, and $\mathrm{BC}_\alpha(t)=\log(t)$ if $\alpha=0$. Liu et al. [7] has demonstrated that such a model can accommodate various different crossing patterns of the hazard rate functions. All the proposed methods discussed in this paper can be adapted easily to such a generalized setting. Model (2) can also be generalized to include the quadratic term of $M(t)$ and other terms in the exponential part on its right-hand-side. Second, model (2) can be generalized easily to include more than one crossing point. Third, the covariates in $\boldsymbol{X}$ and those in $\boldsymbol{W}$ can have some variables in common. See the numerical example of Table 2 in the next section for a demonstration. Fourth, models (1) and (2) can be estimated by maximizing the likelihood of the observed survival and longitudinal data, as in a regular joint modeling approach (cf., [9]). The EM algorithm and the Newton-Raphson algorithm can be used for obtaining parameter estimates. In the model estimation, the term $\phi\,(t-\gamma)\,g$ on the right-hand-side of model (2) can be written as

$$\phi\,(t-\gamma)\,g=\phi tg-\phi\gamma g=:\phi tg-\xi g,$$

where $\xi=\phi\gamma$. By this re-parameterization, the original parameters (ϕ,γ) can be replaced by (ϕ,ξ), and the exponential component of model (2) is linear with respect to both ϕ and ξ, which simplifies the model estimation. In cases when the estimate of ϕ is not significantly different from 0 or when the estimate of γ is outside the time interval $[\min(O_i,i=1,2,,\ldots,n),\max(O_i,i=1,2,,\ldots,n)]$, we conclude that there does not exist any crossing point, although the estimation algorithm still converges.

2.2. Model selection

In practice, there could be many covariates involved in models (1) and (2). It will increase the variability of the estimated model dramatically if an irrelevant covariate is included in a selected model. Therefore, proper selection of a final model is important to increase the efficiency of the related statistical inferences. In the literature on joint modeling of survival and longitudinal data, existing research on model selection is limited. In this subsection, we discuss four criteria for assessing model fit that have been used in the model selection literature, and apply them to the current joint modeling problem with crossing hazard rate functions.

The joint likelihood function of models (1) and (2) can be defined by

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} L_i\left(\boldsymbol{\theta}|O_i, \Delta_i, \boldsymbol{Y_i}, \boldsymbol{t_i}\right) \\ &= \prod_{i=1}^{n}\left[\int_{-\infty}^{\infty}\left\{\prod_{j=1}^{n_i} f_1\left(Y(t_{ij})|\boldsymbol{b}_i, \boldsymbol{\theta}_y\right)\right\} f_2\left(\boldsymbol{b}_i|\boldsymbol{\theta}_b\right) f_3\left(O_i, \Delta_i|\boldsymbol{b}_i, \boldsymbol{\theta}_t\right) d\boldsymbol{b}_i\right], \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_y, \boldsymbol{\theta}_b, \boldsymbol{\theta}_t)$ is a vector of all parameters in models (1) and (2), $\boldsymbol{\theta}_y = \left(\boldsymbol{\beta}, \sigma_\varepsilon^2\right)$, $\boldsymbol{\theta}_b$ is a vector of all parameters in Σ_b of the random effects term in (1), $\boldsymbol{\theta}_t = (\psi, \phi, \gamma, \Lambda_0)$ includes all parameters in model (2) in which $\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$ is the cumulative baseline hazard, $\boldsymbol{Y}_i = (Y(t_{i1}), Y(t_{i2}), \ldots, Y(t_{in_i}))'$, $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \ldots, t_{in_i})'$,

$$f_1\left(Y(t_{ij})|\boldsymbol{b}_i, \boldsymbol{\theta}_y\right) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left\{-\frac{\left(Y(t_{ij}) - M(t_{ij})\right)^2}{2\sigma_\epsilon^2}\right\},$$

$$f_2\left(\boldsymbol{b}_i|\boldsymbol{\theta}_b\right) = \frac{1}{(2\pi)^{q/2}\left|\Sigma_b\right|^{1/2}} \exp\left\{-\frac{1}{2}\boldsymbol{b}_i'\Sigma_b^{-1}\boldsymbol{b}_i\right\},$$

and

$$\begin{aligned} f_3\left(O_i, \Delta_i|\boldsymbol{b}_i, \boldsymbol{\theta}_t\right) &= \left[\lambda_0(O_i)\exp\left\{\psi M(O_i) + \phi\left(O_i - \gamma\right) g_i\right\}\right]^{\Delta_i} \\ &\times \exp\left[-\int_0^{O_i} \lambda_0(u)\exp\left\{\psi M(u) + \phi\left(u - \gamma\right) g_i\right\} du\right]. \end{aligned}$$

The Akaike Information Criterion (AIC), originally introduced by Akaike [22], is a popular model assessment criterion for model selection. Its AIC score is defined by

$$\text{AIC} = -2\log(L_{max}) + 2k,$$

where L_{max} is the maximized value of the likelihood function of the current model under consideration, and k is the number of parameters in the model. By this criterion, among all candidate models, the one with the smallest AIC value is selected.

In models (1) and (2), there are quite a few parameters involved, and the percentage of censored observations is usually quite high in the observed data. In such cases, the corrected AIC (AICc) criterion by Hurvich and Tsai [23] should be relevant, which is designed specifically for cases when the sample size is small or when the number of parameters is a moderate to large fraction of the sample size. The AICc score of this criterion is defined by

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1}.$$

Obviously, the AICc score adds some extra penalty to the AIC score for cases when k is large and $n - k - 1$ is small.

In the model selection literature, the Bayesian Information Criterion (BIC) is another popular model assessment criterion for model selection. The BIC score of this criterion is defined by

$$\text{BIC} = -2\log(L_{max}) + k\log(n).$$

By comparing the BIC score with the AIC score, it can be seen that the former puts more penalty on the number of parameters k in cases when $\log(n) > 2$. As well discussed in the literature, the BIC criterion has the consistency property in the sense that the probability of selecting the true model would approach 1 when n increases and when the true model is among all candidate models under consideration, and the AIC criterion has the asymptotic optimality property in the sense that it can asymptotically choose the best possible model in cases when the true model is not a candidate model (cf., [24], [25]). For analyzing the censored survival data, Volinsky and Raftery [26] proposed a corrected BIC (BICc) criterion with the BICc score

$$\text{BICc} = -2\log(L_{max}) + k\log(r),$$

where r is the number of uncensored survival times.

To use one of the above criteria for model selection, if the number of possible models is relatively small, then we can use the all-subset variable selection procedure. By this procedure, we estimate all possible models and choose the one with the smallest value of the selected model assessment criterion as our final model. If the number of possible models is relatively large, then the backward, forward, stepwise or other variable selection procedures can be considered. See a textbook, such as [27], for a comprehensive description of these variable selection procedures.

2.3. Model diagnostics

After obtaining a final joint model using the model selection method described in the previous subsection, we need to check the model goodness-of-fit and make some appropriate model diagnostics, which is discussed in this subsection. Because the joint model consists of two parts (i.e., the survival part and the longitudinal part) and the responses in these two parts may not be comparable, our proposed model diagnostics are separated for the two parts as well.

For checking the goodness-of-fit of a proportional hazard (PH) model, there are a number of existing methods, including the ones discussed by [28], [29], [30], [31], and [32]. For that purpose, Grønnesby and Borgan [33] proposed a hypothesis test that is similar to the Hosmer-Lemeshow test commonly used in the logistic regression. Their test is based on the martingale residuals after the n subjects are partitioned into m groups according to the estimated risk scores. For the

current PH model (2), the martingale residual of the i^{th} subject is defined as $\Delta_i - r_i$, for $i = 1, 2, \ldots, n$, where

$$r_i = \int_0^{O_i} \widehat{\lambda_0}(u) \exp\left\{\widehat{\psi}\widehat{M}(u) + \widehat{\phi}\,(u - \widehat{\gamma})\, g_i\right\} du, \tag{3}$$

and $\widehat{\lambda_0}(t)$, $\widehat{\psi}$, $\widehat{M}(t)$, $\widehat{\phi}$, and $\widehat{\gamma}$ are estimators of $\lambda_0(t)$, ψ, $M(t)$, ϕ, and γ, respectively. Then, the sum of all martingale residuals within a group would be a random variable with mean 0 if the PH model is valid. The test statistic of the goodness-of-fit test is then defined by

$$T = (H_1, H_2, \ldots, H_{m-1})\widehat{\Sigma}_H^{-1}(H_1, H_2, \ldots, H_{m-1})', \tag{4}$$

where H_j denotes the sum of all martingale residuals within the j^{th} group, for $j = 1, 2, \ldots, m-1$, and $\widehat{\Sigma}_H^{-1}$ is the estimated covariance matrix of $(H_1, H_2, \ldots, H_{m-1})'$. Note that the sum of the martingale residuals of the m^{th} group is not included in (4) because the inverse matrix $\widehat{\Sigma}_H^{-1}$ would not exist otherwise.

As pointed out by May and Hosmer [34], the computation of the inverse matrix $\widehat{\Sigma}_H^{-1}$ in (4) is complicated, and thus the testing procedure based on T is difficult to compute. To overcome this difficulty, May and Hosmer proposed an equivalent but simpler procedure by adding $m-1$ group indicator variables to the PH model and by testing whether the coefficients of all these indicator variables are significantly different from zero, using the score test or the asymptotically equivalent likelihood ratio test. However, both the procedure (4) by Grønnesby and Borgan [33] and its simplified version by May and Hosmer [34] cannot be applied to the current joint modeling problem directly for the following reason. In their models, there are no time-dependent covariates involved in the risk scores, and thus the n subjects can be partitioned by the estimated risk scores in such cases. As a comparison, in model (2) of the current joint modeling problem, both $M(t)$ and $(t-\gamma)g_i$ on the right-hand-side of the model are time-dependent. In such cases, it is difficult to properly partition the n subjects into m groups based on the estimated risk scores because they have the time-dependent covariates involved as well. If we partition the subjects based on the estimated risk scores using subject-specific times, then the results would be biased.

In this paper, we suggest the following procedure to check the goodness-of-fit of the survival part of the joint model. First, we partition all n subjects into m groups by the baseline values of the response variable $Y(t)$ (i.e., by $Y(0)$). Then, $m-1$ group indicator variables are added to model (2), and the likelihood ratio test is performed for testing whether all coefficients of these indicator variables are 0. Regarding the selection of m, one good rule of thumb suggests selecting m to be a simple integer close to the integer part of $2n^{2/5}$ (cf., [35]), and on average each group should include at least 5 subjects. For the determination of the m groups, one well-known practical guideline is that the groups should be chosen in a way such that the numbers of subjects are roughly the same among different groups.

The quantities $\{r_i, i = 1, 2, \ldots, n\}$ defined in (3) can also be used for making a diagnostic plot described as follows. From (3), it is obvious that

$$r_i = \widehat{\Lambda}(O_i) = -\log\left(\widehat{S}(O_i)\right) = -\log\left(1 - \widehat{F}(O_i)\right), \tag{5}$$

where $\widehat{\Lambda}(t)$, $\widehat{S}(t)$ and $\widehat{F}(t)$ are estimators of the cumulative hazard function $\Lambda(t)$, the survival function $S(t)$, and the cumulative distribution function $F(t)$, respectively, defined by (3) and the relationships that $\widehat{\Lambda}(t) = -\log(\widehat{S}(t))$ and $\widehat{S}(t) = 1 - \widehat{F}(t)$. Therefore, If the PH model (2) fits the data well, then roughly speaking the cumulative distribution function of the quantity O_i would be close to $\widehat{F}(t)$. Thus, the distributions of both $\widehat{F}(O_i)$ and $1 - \widehat{F}(O_i)$ would be close to $Uniform[0, 1]$. Consequently, the distribution of r_i is close to $Exponential(1)$, according to (5). For the $Exponential(1)$ distribution, we have

$$S(t) = \exp(-t) \implies -\log(S(t)) = t.$$

Therefore, if the PH model (2) fits the data well, then the scatter plot of $\{(r_i, -\log(S(r_i)), i = 1, 2, \ldots, n\}$ should wave around a straight line through the origin with a slope of 1.

For checking the goodness-of-fit of the longitudinal model (1), there are some existing methods for model diagnostics in the longitudinal data analysis literature. However, as pointed out by Rizopoulos et al. [36], these methods cannot be applied to the joint modeling problem directly, because there are many nonrandom dropouts in the longitudinal outcome due to the occurrence of events in the survival data (see also [37]). More specifically, in cases when individuals in the study experience the event of interest (e.g., death), they will leave the study after the occurrence of the event. In such cases, it is impossible to collect further longitudinal observations after the event times. If we apply the conventional model diagnostics methods designed for ordinary longitudinal data analysis to such data and treat the unavailable longitudinal observations as missing values, then the results could be misleading because the unavailable longitudinal observations are not missing at random. To overcome this difficulty, Rizopoulos et al. [36] proposed a multiple imputation (MI) procedure for generating multiple sets of completed data. This procedure consists of 4 steps briefly described below.

Step 1. Draw an observation of $\boldsymbol{\beta}$ in random from the distribution $N(\widehat{\boldsymbol{\beta}}, \widehat{\Sigma}_{\boldsymbol{\beta}})$, where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\Sigma}_{\boldsymbol{\beta}}$ are the estimated value of $\boldsymbol{\beta}$ and the estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$ obtained in the joint model estimation.

Step 2. Draw an observation of $\boldsymbol{b}_i$ in random from its posterior distribution given the observed data and the coefficient vector $\boldsymbol{\beta}$ generated in Step 1.

Step 3. For a missing longitudinal observation at a given time point t^*, generate its imputed value in random from the distribution

$$N\left(\boldsymbol{X}(t^*)'\boldsymbol{\beta} + \boldsymbol{Z}(t^*)'\boldsymbol{b}_i, \widehat{\sigma}_\varepsilon^2\right),$$

where $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ are generated in Steps 1 and 2, and $\widehat{\sigma}_\varepsilon^2$ is obtained in the joint model estimation.

Step 4. Repeat Steps 1–3 L times for each subject, where L denotes the number of multiple imputations.

In the above procedure, for each subject, to determine the time points for imputation (i.e., t^*), we first find the time point τ which is the maximum of $\{O_i, i = 1, 2, \ldots, n\}$. Then, for the i^{th} subject, we impute its Y values at all time points in $[O_i, \tau)$ that at least one of the other subjects has observations of Y. Since the MI procedure described above requires a complete specification of the likelihood function, the nonparametric baseline hazard function $\lambda_0(t)$ in our PH model (2) should be replaced by a parametric one in order to use this approach. In all our numerical examples discussed in the next two sections, $\lambda_0(t)$ is modeled by a piecewise-constant function. After the imputation step, we can use the traditional model diagnostic approaches as usual on the observed residuals and the residuals generated by the MI procedure (called MI residuals hereafter). In this paper, we use the subject-specific residuals for checking the homoscedasticity and the normality assumptions of the longitudinal model (1). The subject-specific residuals are defined by

$$\widehat{\varepsilon}_{ij} = Y(t_{ij}) - \boldsymbol{X}(t_{ij})'\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}(t_{ij})'\widehat{\boldsymbol{b}}_i.$$

They measure the deviation of the observed values of Y from their predicted values based on the longitudinal model (1). See, for instance, [38] for more discussion on subject-specific residuals.

3. Simulation Study

In this section, we present some simulation results to evaluate the numerical performance of the proposed methods described in the previous section. In the first example, we assume that $\sigma_\varepsilon^2 = 1$ in the longitudinal model (1) and

$$M(t) = \beta_0 + \beta_1 t + \beta_2 g + \beta_3 x_1 + \beta_4 x_2 + \beta_5 x_3 + bt,$$

where g (equals 0 or 1) is a group indicator, x_1, x_2 and x_3 are three time-independent covariates with values generated from $N(\mathbf{0}, I_{3\times 3})$, b is the random effects term having the distribution $N(0, \sigma_b^2)$ with $\sigma_b = 0.1$, and the regression coefficients take the values $\beta_0 = 0$, $\beta_1 = 0.5$, $\beta_2 = -0.3$, $\beta_3 = 0.4$, $\beta_4 = 0$, and $\beta_5 = 0$. The true survival model (2) is assumed to be

$$\lambda(t) = \lambda_0(t) \exp\left[\psi M(t) + \phi\,(t-\gamma)\,g + \eta_1 x_1 + \eta_2 x_2 + \eta_3 x_3\right],$$

where $\lambda_0(t) \equiv 0.05$, $\psi = 0.5$, $\phi = -0.2$, $\gamma = 4$, $\eta_1 = 0$, $\eta_2 = 0.6$, and $\eta_3 = 0$. In such cases, the log ratio of the hazard rate functions of the two groups is shown in Figure 1, which is a straight line with a negative slope of -0.2. Because the log ratio passes 0 at $t = 4$, the two hazard rate functions cross each other at $t = 4$.

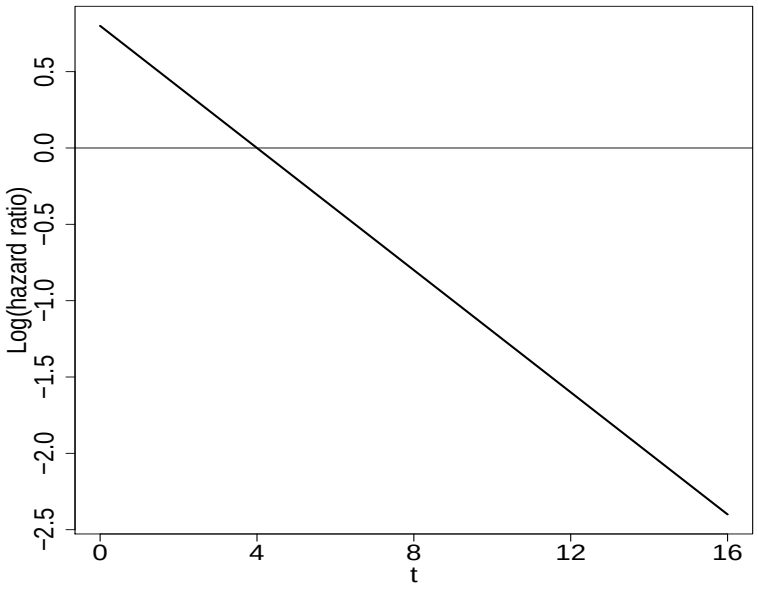

Figure 2. The log ratio of the two hazard rate functions considered in the first simulation example.

In the simulation study, the sample sizes considered are $n = 100, 200$, and 400, with the two treatment groups having the same number of subjects. For each subject, the survival time is generated by the survival model specified above, with the censoring time following the $Uniform[0, 20]$ distribution. In such cases, the censoring rate is about 35%. The longitudinal data are generated at consecutive discrete times with a step of 0.5 in the interval $[0, \min(50, O))$, where O is the observed survival time of a subject. The number 50 is chosen here because the maximum value of $\{O_i, i = 1, 2, \ldots, n\}$ is about 50 in the cases considered.

Then, we use the all-subset model selection procedure to select a final joint model with one of the four model assessment criteria AIC, AICc, BIC, and BICc. In the longitudinal model, we assume that the terms β_0, $\beta_1 t$ and bt are always included in the model. In the survival model, we assume that the term $\psi M(t)$ is always in the model. All other terms in the two models may or may not be included. Therefore, there are a total of $2^4 \times 2^4 = 256$ possible joint models. For each model assessment criterion and each sample size, we record the proportion of times when the true joint model is chosen and the proportion of times when a joint model with the crossing term $\phi\,(t - \gamma)\,g$ included in the survival model (note: such a model may not be the correct model if other terms are not chosen properly) is chosen, among 100 replicated simulations. The second proportion is recorded in this example because it can tell us how well each model assessment criterion can help us choose a model with the correct crossing pattern. Obviously, the second proportion is always larger than or equal

Table 1. For each model assessment criterion and each sample size, this table presents the proportion of times when the true joint model is chosen (Part I), the proportion of times when a joint model with the crossing term included in the survival model is chosen (Part II), and the proportion of times when the true model or one of the six alternative models listed in Table 2 is chosen (Part III). The results are based on 100 replicated simulations.

n	Part I				Part II				Part III			
	AIC	AICc	BIC	BICc	AIC	AICc	BIC	BICc	AIC	AICc	BIC	BICc
100	0.40	0.41	0.26	0.31	0.78	0.70	0.33	0.39	0.69	0.63	0.29	0.35
200	0.50	0.55	0.64	0.63	0.98	0.97	0.74	0.77	0.94	0.94	0.73	0.76
400	0.57	0.59	0.95	0.93	1.00	1.00	0.99	0.99	1.00	1.00	0.99	0.99

to the first proportion. The results are presented in Parts I and II of Table 1. From the table, it can be seen that the four model assessment criteria could lead to different model choices. As expected, BIC and BICc seem more powerful to select the true model than AIC and AICc when the sample size is reasonably large (i.e., $n \geq 200$). When the sample size is small, AIC and AICc have better performance to choose the true model. Also, regarding their performance to choose the true model, AICc is slightly better than AIC, and BIC is slightly better than BICc when $n \geq 200$. Regarding their performance to choose joint models with the correct crossing term, it seems that AIC and AICc perform better than BIC and BICc, especially when the sample size is small to moderate. When the sample size is large (i.e., $n \geq 400$ in the table), their performance is comparable. Based on this example, it seems that the AIC and AICc model assessment criteria should be used if we are quite sure that a crossing term should be included in the survival model, which can be judged based on the plot of the two life-table estimates of the hazard rate functions (cf., Figure 1).

From Table 1, it seems that the four model assessment criteria have quite large chances to choose wrong models. If we check their selected models carefully, then it can be found that most of these models are actually quite close to the true model. For instance, in cases when $n = 200$, the true model and the top six models that are chosen most often based on the four model assessment criteria are listed in Table 2. Because the true model is nested in all these six models, the likelihood ratio test (LRT) can be carried out to compare the true model with each of these six models. Based on 100 replications, the averaged p-values of the six LRT tests are listed in the last column of Table 2. It can be seen that all these averaged p-values are larger than or equal to 0.45, implying that the six alternative models are not significantly different from the true model. If we treat these six alternative models as true models and re-count the proportion of times when the true models are chosen in 100 replicated simulations, the results are shown in Part III of Table 1. From the table, it can be seen that both AIC and AICc perform really well, and BIC and BICc perform well when the sample size is moderate to large (i.e., $n = 200$ or 400 in the table).

Next, we investigate our proposed goodness-of-fit test described in the paragraph immediately before the one containing

Table 2. Averaged p-values of the LRT tests to compare the true model with the top six models that are chosen most often based on the four model assessment criteria in cases when $n = 200$. The results are based on 100 replications.

Models	Covariates		Averaged p-values
	Longitudinal part	Survival part	
True model	g, x_1	g, x_2	
Model 1	g, x_1, x_3	g, x_1, x_2	0.52
Model 2	g, x_1	g, x_1, x_2	0.54
Model 3	g, x_1, x_3	g, x_2, x_3	0.46
Model 4	g, x_1	g, x_2, x_3	0.45
Model 5	g, x_1, x_2	g, x_2	0.46
Model 6	g, x_1, x_3	g, x_2	0.50

expression (5). The survival model considered is the true one described at the beginning of this section in the case when $n = 200$. In the test, all subjects are divided into 10 groups based on the values of $Y(0)$ in the way as described in Subsection 2.3. The simulation is repeated 100 times, generating 100 p-values of the test. The overall estimate of the p-value is defined by their average, and the standard error (SE) of the overall estimate is defined by their sample standard deviation divided by $\sqrt{100} = 10$, which are 0.508 and 0.028, respectively. Because the estimated p-value is large, the test concludes that the survival model fits the data well, which confirms that our proposed goodness-of-fit test performs well in this example. As a comparison, we also try the test that is exactly the same as our proposed one, except that the subjects are randomly divided into 10 groups. The corresponding overall estimate of the p-value and the standard error are 0.458 and 0.028, respectively. It can be seen that this alternative test also performs well, although the first one is a little better.

At the end of this section, we consider another example to investigate whether the longitudinal modeling (cf., expression (1)) is helpful for estimating the survival model (2) and the related crossing point. For this purpose, let us assume that the true survival model is

$$\lambda(t) = \lambda_0(t) \exp\{\psi M(t) + \phi(t - \gamma) g\}, \tag{6}$$

where $\psi = 0.5, \phi = 0.2, \gamma = 4$,

$$M(t) = \beta_0 + (\beta_1 + b) t,$$

$\beta_0 = 1$, $\beta_1 = 0.5$, and b is the random effect coefficient generated from a zero-mean normal distribution with variance $\sigma_b^2 = 0.1$. The observed longitudinal data (i.e., observations of $Y(t)$) are generated from $N(M(t), \sigma_\varepsilon^2)$ with $\sigma_\varepsilon^2 = 0.5$. The sample sizes considered are $n = 200$ and 400, with the two treatment groups having the same number of subjects. The censoring time for each subject is generated from the $Uniform[1, T]$ distribution, where T is adjusted to reach a pre-specified censoring rate. We choose the left end of the interval to be 1 to allow a minimum follow-up time of one time unit. In the simulation, we consider the two censoring rates of 20% and 50%. Besides the joint modeling approach described

in Subsection 2.1, we also consider the so-called "naive" approach, by which $M(t)$ in (6) is replaced by $Y(t)$. Namely, we do not specifically model the observed longitudinal data in the naive approach. For both methods, the simulation is repeated for 100 times, and the MSE values of their estimators of ψ, ϕ, and γ are computed and presented in Table 3. From the table, it can be seen that the joint modeling approach is more effective than the naive approach in estimating the three parameters of model (6) in all cases, except that the two approaches are comparable in estimating γ when $n = 200$ and the censoring rate is 50%. This example demonstrates that proper modeling of the longitudinal data can generally improve the estimation of the survival model.

Table 3. MSE values of the estimators of ψ, ϕ, and γ in model (6) of the joint modeling and naive approaches based on 100 replicated simulations.

n	censoring rate	Method	ψ	ϕ	γ
200	20%	joint modeling	0.0750	0.0086	1.7837
		naive	0.1643	0.0099	1.8693
	50%	joint modeling	0.1471	0.0335	2.1464
		naive	0.1803	0.0353	2.1392
400	20%	joint modeling	0.0538	0.0037	0.5528
		naive	0.1332	0.0041	0.7360
	50%	joint modeling	0.1198	0.0114	1.6661
		naive	0.1903	0.0133	2.0055

4. A Real-Data Example

The early breast cancer data described briefly in Section 1 were obtained from a clinical trial study designed to assess the time to death in 231 patients with early breast cancer. Those 231 patients were randomly assigned to receive either the Cyclophosphamide Epirubicin Fluorouracil (CEF) treatment or the Cyclophosphamide Methotrexate Fluorouracil (CMF) treatment after their breast cancer surgeries, among which 107 patients received the CEF treatment and the remaining 124 patients received the CMF treatment. For each patient, her survival time was defined to be the period from the start of the study to the time when she died. Besides the survival data, the following longitudinal data were also available. During the course of the study, observations of a numerical index of quality of life (QOL) were recorded for each patient at times when she visited the clinic. Furthermore, there are four time-independent covariates involved, described below.

trt: treatment group indicator: 1 = CEF, 0 = CMF,

age: age of each patient at the beginning of study,

node: number of axillary nodes that each patient had at the beginning of study,

size: size of tumor that each patient had at the beginning of study.

The life-table estimates of the two hazard rate functions are shown in Figure 1 and it shows that the two hazard functions corresponding to the CEF and CMF treatment groups cross each other at around $t = 60$ months. Thus, we might conclude that the treatment effects of CEF and CMF are different in different stages of the early breast cancer, and the survival model with a single crossing point is reasonable to use. For the longitudinal data, we consider the following full model:

$$Y(t_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{trt}_i + \beta_3 \text{age}_i + \beta_4 \text{node}_i + \beta_5 \text{size}_i + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}, \tag{7}$$

where Y is the observed index of QOL, $\mathbf{b}_i = (b_{0i}, b_{1i})'$ are the random effects coefficients following the distribution $N(0, \Sigma_b)$, $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ are regression coefficients, and ε_{ij} are i.i.d. random errors. For the survival data, we consider the following Cox's PH model with one crossing point:

$$\lambda(t_{ij}) = \lambda_0(t_{ij}) \exp\left\{\psi M(t_{ij}) + \phi\,(t_{ij} - \gamma)\,\text{trt}_i + \eta_1 \text{age}_i + \eta_2 \text{node}_i + \eta_3 \text{size}_i\right\}, \tag{8}$$

where $M(t_{ij})$ is the right-hand-side of equation (7) without the random error term, γ is the crossing point, and $(\psi, \phi, \eta_1, \eta_2, \eta_3)$ are coefficients. Since there are four covariates (i.e., trt, age, node, size) involved in both the longitudinal and survival models (7) and (8), there are a total of $2^4 \times 2^4 = 256$ possible models to consider.

Then, we use the all-subset model selection procedure to choose our final joint model. As pointed out in the previous section, if we are quite certain about the crossing phenomenon of the two hazard rate functions, then the model selection results based on AIC and AICc would be more reliable than those based on BIC and BICc. Since Figure 1 already demonstrates an obvious crossing pattern, the AIC and AICc model assessment criteria are considered here. Under both criteria, the selected final joint model consists of the longitudinal model

$$Y(t_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{trt}_i + \beta_3 \text{age}_i + \beta_4 \text{node}_i + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}, \tag{9}$$

and the survival model

$$\lambda(t_{ij}) = \lambda_0(t_{ij}) \exp\left\{\psi M(t_{ij}) + \phi\,(t_{ij} - \gamma)\,\text{trt}_i\right\}. \tag{10}$$

By the way, if we restrict the final joint model to contain the crossing term $\phi(t_{ij} - \gamma)\text{trt}_i$ in its survival model, then the selected final joint models based on the BIC and BICc model assessment criteria are exactly the same as those described by (9) and (10). If we do not add that restriction, then their selected final joint models do not contain the crossing term in the survival models.

The estimated parameters and the corresponding 95% asymptotic confidence intervals (CIs) of the parameters in models (9) and (10) are presented in Table 4. Remember that the parameter γ can be expressed as $\gamma = \xi/\phi$, as discussed at the end

Table 4. Point estimates and 95% asymptotic CIs of the parameters in models (9) and (10).

	Parameters	Estimates	95% CIs
Model (10)	ψ	0.2655	$(0.2359, 0.2952)$
	ϕ	0.0144	$(0.0047, 0.0241)$
	γ	56.7314	$(38.1224, 75.3404)$
Model (9)	β_0	4.0802	$(4.0280, 4.1324)$
	β_1	0.0237	$(0.0233, 0.0241)$
	β_2	-0.1248	$(-0.1410, -0.1086)$
	β_3	0.0239	$(0.0228, 0.0251)$
	β_4	0.0101	$(0.0084, 0.0118)$

of Subsection 2.1 about the model re-parameterization. So, its standard error can be computed by the delta-method from the estimated covariance matrix of the estimator of (ϕ, ξ). From the table, it can be seen that the estimate of ψ is 0.2655 in the survival model (10) and its 95% asymptotic confidence interval does not contain 0. Thus, we could conclude that there is a significant evidence to support the statement that QOL has a positive association with the hazard of death. This is really an interesting result because it might conflicts with our intuition. However, if we think about the related variables carefully, then we would agree that women with a high social status are often more likely to have a good QOL, and they usually have more stress as well. It is well known that stress is a main risk factor of breast cancer (cf., [39]). Therefore, it is possible that women with high QOL would have more chance to get breast cancer.

The ratio of the two hazard rate functions of the CEF and CMF treatment groups is

$$\begin{aligned} & \frac{\lambda(t_{ij}|\text{trt}_i = 1)}{\lambda(t_{ij}|\text{trt}_i = 0)} \\ = \; & \frac{\lambda_0(t_{ij}) \exp\left\{\psi\left(\beta_0 + \beta_1 t_{ij} + \beta_2 + \beta_3 \text{age}_i + \beta_4 \text{node}_i + b_{0i} + b_{1i} t_{ij}\right) + \phi\left(t_{ij} - \gamma\right)\right\}}{\lambda_0(t_{ij}) \exp\left\{\psi\left(\beta_0 + \beta_1 t_{ij} + \beta_3 \text{age}_i + \beta_4 \text{node}_i + b_{0i} + b_{1i} t_{ij}\right)\right\}} \\ = \; & \exp\left\{\psi\beta_2 + \phi\left(t_{ij} - \gamma\right)\right\}. \end{aligned} \tag{11}$$

From equation (11), it can be easily calculated that the two hazard rate functions cross each other at $\gamma - \frac{\psi}{\phi}\beta_2$. Note that this number is different from γ because $M(t_{ij})$ in (10) depends on trt_i (cf., (9)) and γ can be interpreted as the crossing point only when the value of trt_i is given. By replacing the related parameters with their estimates shown in Table 4, the estimated crossing point is 59.0324. So, it seems that the hazard rate of the CMF treatment group is higher than the hazard rate of the CEF treatment group before $t = 59.0324$ months, and their relative positions are switched after that time point. The estimated log hazard ratio by our joint modeling approach is shown by the dashed line in Figure 3, and the estimated log hazard ratio based on the life-table estimates of the two hazard rate functions is shown in the same plot by the solid curve. From the plot, it can be seen that the estimate by the joint modeling approach matches the empirical estimate based on the survival data alone well, and their estimated crossing points are close to each other.

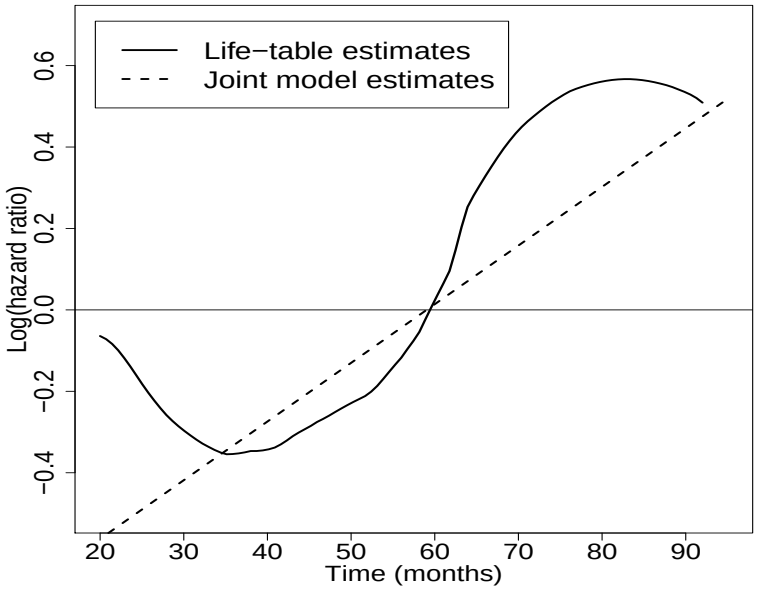

Figure 3. Estimated log hazard ratio by our joint modeling approach (dashed line) and the empirical estimate based on the life-table estimates of the two hazard rate functions (solid curve) of the CEF and CMF treatment groups in the early breast cancer example.

For the longitudinal model (9), from Table 4, it can be seen that all parameters β_0, β_1, β_2, β_3, and β_4 are significantly different from 0 at the 0.05 significance level, because all their 95% CIs do not cover 0. The positive value of $\widehat{\beta_1} = 0.0237$ suggests that QOL increases over time. This may reflect the fact that most patients were hard to accept the fact that they had the cancer at the beginning of study, and they accepted this fact gradually and adjusted their lifes properly over time. The negative value of $\widehat{\beta_2} = -0.1248$ shows that the patients receiving the CMF treatment had higher QOL scores than those receiving the CEF treatment. Also, the QOL scores seem to increase with age, and it has a small but positive association with the number of axillary nodes.

Next, we check the goodness-of-fit of the selected joint model. For the estimated survival model (10), we use the baseline QOL scores of patients (i.e., the QOL scores at the beginning of study) to partition all patients into 15 groups, where the number 15 is chosen to be a simple integer number that is close to the number $2n^{2/5} = 17.639$. The groups are determined by the $(j/15)$-th quantiles of the baseline QOL scores, for $j = 1, 2, \ldots, 15$. The overall goodness-of-fit test is performed by using the likelihood ratio test after adding 14 group indicators to the model (10). The resulting p-value of the test is 0.570, implying that the joint model fits the data well. For the 15 groups of patients, the observed and expected numbers of events are computed and presented in Table 5. For each group, we can use the z-test with the test statistic

$$\frac{\text{observed \# of events} - \text{expected \# of events}}{\sqrt{\text{expected \# of events}}},$$

to check whether the difference between the observed and expected numbers of events is significant. The observed z-scores and the corresponding p-values are shown in Table 5 as well. From the table, it can be seen that only one group

Table 5. Observed and expected numbers of events, z-scores, and p-values of the 15 groups of patients in the early breast cancer example.

Group	Observed number of events	Expected number of events	z-score	p-value
1	14	15.06	-0.27	0.39
2	13	9.47	1.15	0.13
3	14	16.59	-0.64	0.26
4	14	11.09	0.87	0.19
5	14	13.37	0.17	0.43
6	16	11.84	1.21	0.11
7	13	18.23	-1.22	0.11
8	15	9.18	1.92	0.03
9	15	16.54	-0.38	0.35
10	15	19.59	-1.04	0.15
11	15	12.39	0.74	0.23
12	14	20.11	-1.36	0.09
13	15	14.29	0.19	0.43
14	16	15.20	0.20	0.42
15	15	14.14	0.23	0.41

(i.e., group 8) has a significant p-value and the remaining 14 groups all have insignificant p-values. This result matches the result of the overall goodness-of-fit test well.

As mentioned in Subsection 2.3, if the estimated model (10) fits the data well, then the scatter plot of $\{(r_i, -\log(S(r_i)), i = 1, 2, \ldots, n\}$ should wave around a straight line through the origin with a slope of 1. For the early breast cancer data, this plot is shown in Figure 4. It can be seen that all points of $\{(r_i, -\log(S(r_i)), i = 1, 2, \ldots, n\}$ are close to the straight line through the origin with a slope of 1. Therefore, this plot confirms that the estimated survival model (10) fits the data well

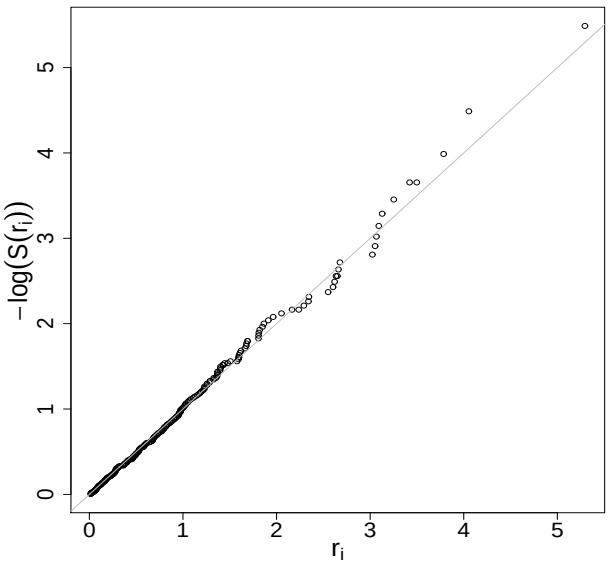

Figure 4. Scatter plot of $\{(r_i, -\log(S(r_i)), i = 1, 2, \ldots, n\}$ in the early breast cancer example. The solid grey line is the reference line through the origin with a slope of 1.

For diagnostics of the estimated longitudinal model (9), the left plot of Figure 5 shows the subject-specific residuals versus the corresponding fitted values of QOL. The grey dashed line in the plot denotes the loess fit of these observed residuals. To handle the nonrandom dropouts, the corresponding unobserved QOL scores are imputed using the 4-step procedure described in Subsection 2.3. The right plot in Figure 5 shows the observed subject-specific residuals (black circles) and the MI subject-specific residuals (grey points). The grey dashed line in the plot is the loess fit of both the observed residuals and the MI residuals. So, the dashed line in the left plot describes the relationship between the observed residuals and the fitted values of QOL while the dashed line in the right plot describes the relationship between the observed/MI residuals and the fitted values. By comparing the two lines, we can see that there is a systematic trend in the former case when the observed residuals are used alone and this trend is mostly eliminated in the latter case.

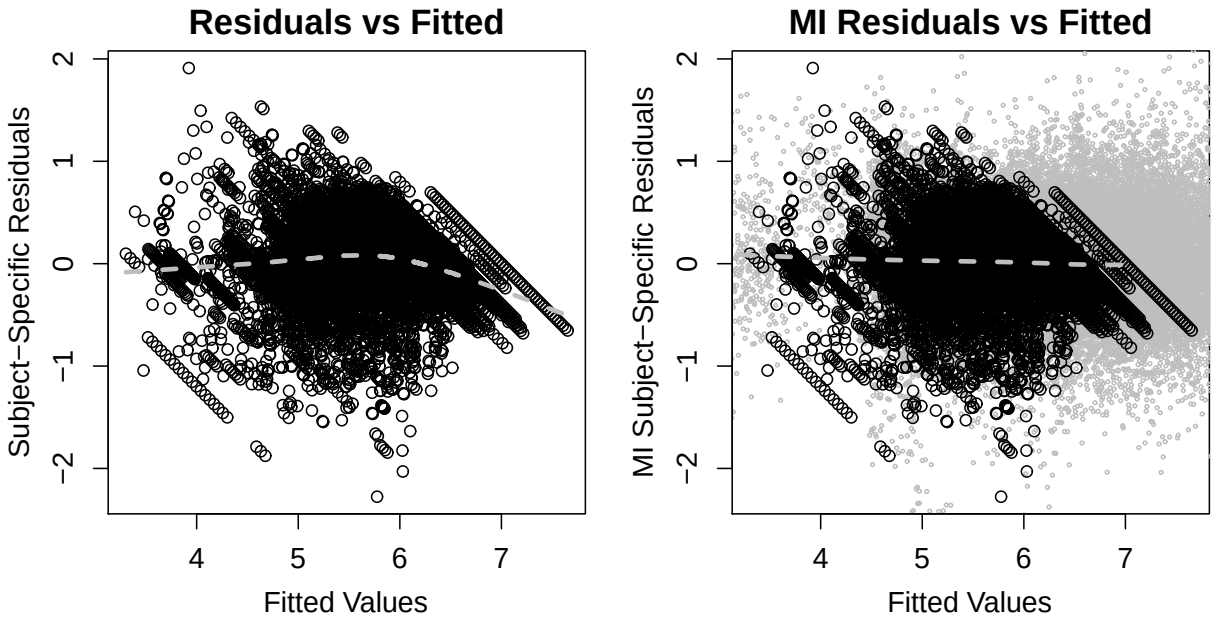

Figure 5. Diagnostic plots of the estimated longitudinal model (8) in the early breast cancer example. The left and right plots show the observed subject-specific residuals versus the fitted values of QOL and the observed and multiple imputed (MI) subject-specific residuals versus the fitted values of QOL, respectively.

5. Concluding Remarks

In the previous sections, we have discussed model estimation, model selection, and model diagnostics for joint modeling the survival and longitudinal data when the hazard rate functions cross each other. For model selection, the four model assessment criteria AIC, AICc, BIC, and BICc are compared using numerical simulations. Based on the numerical study, it seems that AIC and AICc are more reliable for model selection in cases when we are quite sure from the life-table estimates of the hazard rate functions that they cross each other. When the sample size is moderate to large, all four criteria perform

reasonably well. For checking the goodness-of-fit of the survival model, the likelihood ratio test by grouping subjects and a diagnostic plot based on the quantities $\{r_i, i = 1, 2, \ldots, n\}$ are discussed. For checking the goodness-of-fit of the longitudinal model, the 4-step multiple imputation approach is discussed to eliminate the impact of nonrandom dropouts of the longitudinal response.

There are still many issues that need to be addressed in our future research. For instance, when we perform model diagnostics, the methods discussed in the current paper handles the survival and longitudinal models separately. It is still unknown to us how to check the goodness-of-fit of the estimated survival and longitudinal models simultaneously by a single test or by a single diagnostic plot. Furthermore, in grouping subjects when checking the goodness-of-fit of the estimated survival model, we have used the baseline longitudinal response for that purpose and the numerical examples show that it worked well. However, it requires much theoretical research to justify this approach.

Acknowledgments The authors thank the associate editor and two referees for their valuable comments which greatly improved the quality of this paper.

References

1. Lawless JF. *Statistical Models and Methods for Lifetime Data.* New York: John Wiley & Sons, 1982.
2. Bain LJ, Engelhardt M. *Statistical Analysis of Reliability and Life-testing Models: Theory and Methods (2nd edition).* New York: Marcel Dekker, 1991.
3. Klein JP, Moeschberger ML. *Survival Analysis.* New York: Springer-Verlag, 1997.
4. O'Quigley J, Pessione F. Score test for homogeneity of regression effect in the proportional hazards model. *Biometrics* 1989; **45:** 135–144.
5. O'Quigley J, Pessione F. The problem of a covariate-time qualitative interaction in a survival study. *Biometrics* 1991; **47:** 101–115.
6. Lin X, Wang H. A new testing approach for comparing the overall homogeneity of survival curves. *Biometrical Journal* 2004; **46:** 489–496.
7. Liu K, Qiu P, Sheng J. Comparing two crossing hazard rates by Cox proportional hazards modelling. *Statistics in Medicine* 2007; **26:** 375–391.
8. Qiu P, Sheng J. A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society (Series B)* 2008; **70:** 191–208.
9. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997; **53:** 330–339.
10. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 2004; **14:** 809–834.
11. O'Quigley J. On a two-sided test for crossing hazard rates. *The Statistician* 1994; **43:** 563–569.
12. Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP. Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* 1980; **36:**, 607–625.
13. Mantel N, Stablein DM. The crossing hazard function problem. *The Statistician* 1988; **37:** 59–64.
14. Moreau T, Maccario J, Lellouch J, Huber C. Weighted log rank statistics for comparing two distributions. *Biometrika* 1992; **79:** 195–198.
15. Anderson JA, Senthilselvan A. A two-step regression model for hazard functions. *Applied Statistics* 1982; **31:** 44–51.
16. Breslow NE, Edler L, Berger J. A two–sample censored-data rank test for acceleration. *Biometrics* 1984; **40:** 1049–1062.
17. Zhang JJ, Peng YW. Crossing hazard functions in common survival models. *Statistics and Probability Letters* 2009; **79:** 2124–2130.
18. Cheng MY, Qiu P, Tan X, Tu D. Confidence intervals for the first crossing point of two hazard functions. *Lifetime Data Analysis* 2009; **15:** 441–454.
19. Muggeo VMR, Tagliavia M. A flexible approach to the crossing hazards problem. *Statistics in Medicine* 2010; **29:** 1947–1957.
20. Zeng D, Cai J. Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *The Annals of Statistics* 2005; **33:** 2123–2163.
21. Duphy J, Grama I, Mesbah M. Asymptotic theory for the Cox model with missing time-dependent covariate. *The Annals of Statistics* 2006; **34:** 903–924.
22. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19:** 716–723.
23. Hurvich CM, Tsai C. Regression and time series model selection in small samples. *Biometrika* 1989; **76:** 297–307.
24. Shao J. An asymptotic theory for linear model selection (with discussions). *Statistica Sinica* 1997; **7:** 221–242.
25. Yang Y. Can the strengths of AIC and BIC be shared? *Biometrika* 2005; **92:** 937–950.
26. Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. *Biometrics* 2000; **56:** 256–262.
27. Weisberg S. *Applied Linear Regression (3rd edition).* John Wiley& Sons: New York, 2005.
28. Aalen OO. Further results on the non-parametric linear regression model in survival analysis. *Statistics in Medicine*, 1993; **12:** 1569–1588.
29. Fleming TR, Harrington DP. *Counting Processes and SurvivaI Analysis*, John Wiley & Sons: New York, 1991.
30. Hjort NL. Goodness of fit tests in models for life history data based on cumulative hazard rates. *Annals of Statistics*, 1990; **18:** 1221–1258.
31. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 1993; **80:** 557–572.
32. Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 1980; **67:** 145–153.

33. Grønnesby JK, Borgan Ø. A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Analysis* 1996; **2:** 315–328.
34. May S, Hosmer DW. A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Analysis* 1998; **4:** 109–120.
35. D'Agostino RB, Stephens MA. *Goodness-of-Fit Techniques.* New York: Marcel Dekker, 1986.
36. Rizopoulos D, Verbeke G, Molenberghs G. Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics* 2010; **66:** 20–29.
37. Gelman A, Mechelen IV, Verbeke G, Heitjan DF, Meulders M. Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics* 2005; **61:** 74–85.
38. Nobre J, Singer J. Residuals analysis for linear mixed models. *Biometrical Journal* 2007; **6:** 863–875.
39. Peled R, Carmil D, Siboni-Samocha O, Shoham-Vardi I. Breast cancer, psychological distress and life events among young women. *BMC Cancer* 2008; **8:** 245.