Joint Modeling of Efficacy and Laboratory Data in Clinical Trials

Kao-Tai Tsai

Celgene Corporation, Summit, New Jersey, USA

Abstract

Joint modeling of time-to-event data such as survival or time to disease progression incorporating with the longitudinal data has been an active research topic. Software programs are also available in the public domain to perform this kind of data analysis. However, most of the programs are limited to one longitudinal data series and to extend this to multiple longitudinal series is remaining as a challenge. In this research, we estimate the treatment effects on disease progression using joint modeling with multiple data series of laboratory tests. We also estimate the variations of estimates via bootstrap. We compare the results from the existing software programs with respect to their consistency. Data from a recent clinical trial is used to illustrate the proposed approach.

Keywords: Laboratory test data, joint model, clinical trial.

1 Introduction

Clinical trials collect huge amount of data. For trials with long study durations, such as cancer studies, longitudinal repeated measurements on the treatment effect with respect to responses, disease progression, or duration of survival are among some of the commonly collected endpoints. In addition to these endpoints of interest, laboratory test data, safety data in terms of adverse events, concomitant medications taken during the studies, etc. are also been collected repeatedly during the studies. With respect to the laboratory test data, even though NIH, Mayo Clinic, and other institutions had published various laboratory test guides for various diseases with the emphasis of their importance, these data are rarely analyzed in full extend. They are most often being summarized with simple summary tables or with data listings.

The longitudinal data, such as laboratory tests, genetic biomarker, or health outcomes, can be important predictors or surrogates of an event of interest, such as progression-free survival, relapse-free survival, or overall survival. Classical models such as the linear mixed effects model for longitudinal data and the Cox proportional hazards model for time-to-event data do not consider dependencies between them. Alternatively, joint models for longitudinal data and time-to-event data are commonly used that bring these two data types together (simultaneously) into a single model so that one can infer the association between them, and to better assess the effect of a treatment. Due to the rapid development of clinical and genetic biomarkers in clinical trials, joint modeling has gained its popularity in recent years because it can potentially provide more efficient estimate of the treatment effects on the event of interest, more efficient estimate of the treatment effects on the longitudinal data series, more detailed relationship on how hazard of events are affected by longitudinal process in dimension of time, and can potentially provide a better estimate of the overall treatment effect because more aspects of the study data are analyzed together.

However, most of the practices are limited to the inclusion of one longitudinal data series. In some cases with multiple data series, one of them is treated as response with the rest of them are treated as covariates. To extend this for general cases including multiple longitudinal series remains as a challenge, in addition, many commonly used software programs seem to also have this limitation. The objective of this research is to extend the current case and propose a joint modeling with multiple longitudinal processes so that to enhance a better understanding of treatment effect by the inclusion of more data series from the huge amount of data the study had collected.

In this research, we estimate the treatment effects on disease progression using joint modeling with multiple data series of laboratory tests. We also estimate the variations of estimates via bootstrap. We compare the results from the existing software programs in R with respect to their consistency. Data from a recent clinical trial is used to illustrate the proposed approach.

2 Notations and Models

For subject i, $(i = 1, ..., N)$, for the time-to-event process, let T_i^* denote the true event time, C_i be the censoring time. The distributions of T_i^* and C_i are independent, the observed $T_i = \min(T_i^*, C_i)$, $\delta_i = I(T_i^* \leq C_i)$ is the event indicator, with hazard function $\lambda_i(t)$. For the longitudinal process, let $y_{ij}(t)$ denote the value of the longitudinal outcome at time point t_{ij} with $j = 1, \cdots, n_i$.

The joint model assumes a longitudinal process

$$y_i(t) = \mathcal{F}_{1i}(t) + \mathcal{R}_{1i}(t) + \epsilon_i(t), \tag{1}$$

where $\mathcal{F}_{1i}(t)$ is a fixed effect, $\mathcal{R}_{1i}(t)$ is an unobserved random effect, and $\epsilon_i(t)$ is random measurement error. It also assumes an event process, such as survival, with hazard function

$$\lambda_i(t) = \lambda_0(t) \exp\{\mathcal{F}_{2i}(t) + \mathcal{R}_{2i}(t)\}, \tag{2}$$

where $\mathcal{F}_{2i}(t)$ is a fixed effect, $\mathcal{R}_{2i}(t)$ is an unobserved random effect. The random effects are assumed to follow a joint normal distribution, i.e., $(\mathcal{R}_{1i}, \mathcal{R}_{2i}) \sim N(0, \Sigma)$.

Specifically for the longitudinal process, at each time point $t \in \{t_{ij} \mid j = 1, \cdots, n_i\}$,

$$\begin{aligned} y_i(t) &= \mathcal{F}_{1i}(t) + \mathcal{R}_{1i}(t) + \epsilon_i(t) \\ &= x_i'(t)\beta + z_i'(t)b_i + \epsilon_i(t), \end{aligned} \tag{3}$$

where $x_i(t)$ is the design matrix for the fixed effect, β is the vector of the unknown fixed effect parameters, $z_i(t)$ is the design matrix for the random effect, and $b_i \sim N(0, \Sigma)$ is a vector of random effect parameters, and $\epsilon_i(t) \sim N(0, \sigma^2)$ is the measurement error independent of b_i.

To quantify the effect of $\mathcal{F}_{1i}(t) + \mathcal{R}_{1i}(t)$ on the risk of an event. One of the common options is to use a relative risk model of the form (Therneau and Grambsch 2000):

$$\begin{aligned} h_i(t|\mathcal{M}_i(t), \omega_i) &= \lim_{dt \to 0} Pr\{t \le T_i^* < t + dt \mid T_i^* \ge t, \mathcal{M}_i(t), \omega_i\}/dt \\ &= h_0(t) \exp\{\gamma^T \omega_i + \alpha(\mathcal{F}_{1i}(t) + \mathcal{R}_{1i}(t))\} \end{aligned} \tag{4}$$

where $\mathcal{M}_i(t) = \{\mathcal{F}_{1i}(u) + \mathcal{R}_{1i}(u), 0 \le u < t\}$ denotes the history of the true unobserved longitudinal process up to time t, $h_0(t)$ denotes the baseline risk function at time t, and $\omega_i = \mathcal{F}_{2i}(t)$ is a vector of baseline covariates with a corresponding vector of regression coefficients γ, and parameter α quantifies the effect of association between the underlying longitudinal outcome and the hazard of an event.

Note that the baseline hazard function $h(\cdot)$ can be estimate at each time point t, namely $h(t)$. One can also estimate $h(\cdot)$ based on the cumulative information of hazard up to time t, namely $h(\mathcal{C}_i(t))$, where

$$\mathcal{C}_i(t) = \int_0^t \exp\{\gamma^T \omega_i + \alpha(\mathcal{F}_{1i}(s) + \mathcal{R}_{1i}(s))\} ds.$$

Hence,

$$h_i(t|\mathcal{M}_i(t), \omega_i) = h_0(\mathcal{C}_i(t)) \exp\{\gamma^T \omega_i + \alpha(\mathcal{F}_{1i}(t) + \mathcal{R}_{1i}(t))\} \tag{5}$$

If prior knowledge about $h(\cdot)$ is available, the extra specification of h can increase the efficiency of estimation.

Wulfsohn & Tsiatis (1997) used the 2-stage method proposed by Laird & Ware (1982) as another approach. Henderson, et al. (2000) extended Wulfsohn & Tsiatis' approach and proposed the following approach.

With the latent bivariate Gaussian process $\mathcal{R}_i(t) = (\mathcal{R}_{1i}(t), \mathcal{R}_{2i}(t))$, such that the longitudinal process

$$Y_{ij} = x_{1i}(t)'\beta_1 + \mathcal{R}_{1i}(t_{ij}) + e_{ij} \tag{6}$$

with

$$\mathcal{R}_{1i}(t) = V_{1i}(t) + d_{1i}(t)'V_{2i}(t),$$

for

$$V_{1i} \sim N(0, \Sigma_{v1}), \quad V_{2i} \sim N(0, \Sigma_{v2}),$$

and the time-to-event process

$$\lambda_i(t) = H_i(t)\alpha_0 \exp\{x_{2i}(t)'\beta_2 + \mathcal{R}_{2i}(t)\}, \tag{7}$$

Henderson et al., assumed

$$V_{1i}(t) = U_1, \ V_{2i}(t) = U_2, \ d_{1i}(t) = t,$$

and

$$\mathcal{R}_{2i}(t) = \gamma_1 U_1 + \gamma_2 U_2 + \gamma_3(U_1 + U_2 t) + U_3$$

with U_3 being another error term.

Note that if $d_{1i}(t) = t_i$, then $\mathcal{R}_{1i}$ becomes a simple random intercept and slope model. One of the differences between the relative risk model and the 2-stage method is the latter approach allows extra random effects in the time-to-event process in addition to that from the longitudinal process to increase the flexibility of individual effect. In addition, the formal approach uses the MLE and the latter approach uses EM algorithm to estimate the parameters. Both methods only analyzed one longitudinal process.

3 Parameter estimation

The joint likelihood function contribution from the i-th subject can be formulated as

$$\mathcal{L} = \int p(T_i; \delta_i | \mathcal{R}_i; \theta_t, \beta) \times \prod_j p(y_i(t_{ij}) | \mathcal{R}_i, \theta_y) \times p(\mathcal{R}_i, \theta_{\mathcal{R}}) d\mathcal{R}_i, \tag{8}$$

namely,

$$\mathcal{L} = P(\text{event process}) \times P(\text{longitudinal process}) \times P(\text{latent random processes}).$$

However, for the multiple longitudinal series, one can not simply extend Eq (8) to the following equation due to the correlation between the longitudinal series:

$$\mathcal{L} = \int p(T_i; \delta_i | \mathcal{R}_i; \theta_t, \beta) \times \prod_j p(y_i(t_{ij}) | \mathcal{R}_i, \theta_y) \times \prod_j p(z_i(t_{ij}) | \mathcal{R}_i, \theta_z) \times \cdots \times p(\mathcal{R}_i, \theta_{\mathcal{R}}) d\mathcal{R}_i. \quad (9)$$

A possible alternative approach is to consider the general linear mixed effects model which incorporates multiple series of repeated measures as dependent variable such as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} \quad (10)$$

where $\boldsymbol{u} \sim \boldsymbol{N(0, G)},\ \boldsymbol{e} \sim \boldsymbol{N(0, R)}$, and $\boldsymbol{Cov(u, e) = 0}$.

Equation (10) includes parameters in the fixed effects vector $\boldsymbol{\beta}$ and all unknowns in the covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$. The number of parameters to be estimated increases almost exponentially when the number of sequences and the number of repeats increased that can cause substantial computational challenges in convergence.

Even though some researchers assume one sequence as the response and other sequences as covariates, others sum up the values of various sequences and treat it as one sequence, these approaches can be problematic and not using the data fully and efficiently. We therefore propose to analyze the data from event process and data from each longitudinal process separately to estimate the association between these two processes, and combing these associations to estimate the overall association between the multiple series and the event process.

4 Example: Joint Model with Two Longitudinal Processes

In this section, we illustrate the proposed method using a recent clinical data set on hematological disorder. The clinical trial data had sample size about 650 subjects. The efficacy data is time to disease progression (PFS) and the safety data consists of laboratory test data series, specifically, the change from baseline of serum protein level and white blood cells (WBC) for each cycle of the study. These data were not measured at every cycle and data beyond cycle 10 were very sparse, we therefore only keep data from cycles 3, 5, 7, 8, 9, and 10.

Figure 1 shows the Kaplan-Meier curve of PFS for the treatment and control groups. The active treatment clearly shows a significant better PFS advantage comparing with the control group.

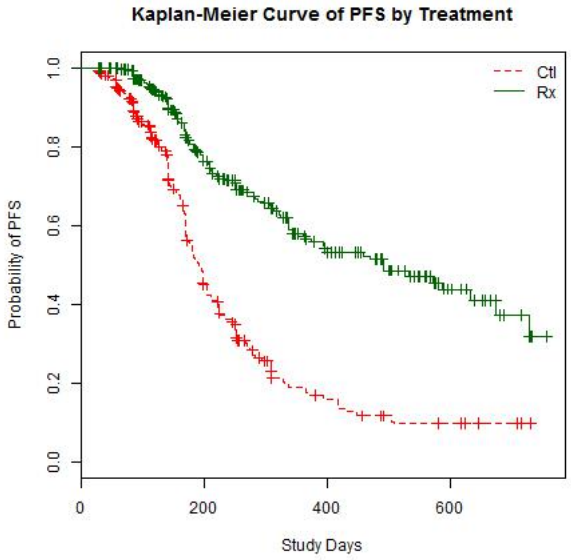

Figure 1: Kaplan-Meier Curves of the Progression-Free-Survival

For the laboratory data, the active treatment had shown an overall larger reduction of serum protein level (Figure 2), while the change of WBC is not as obvious when comparing the active treatment with the control group.

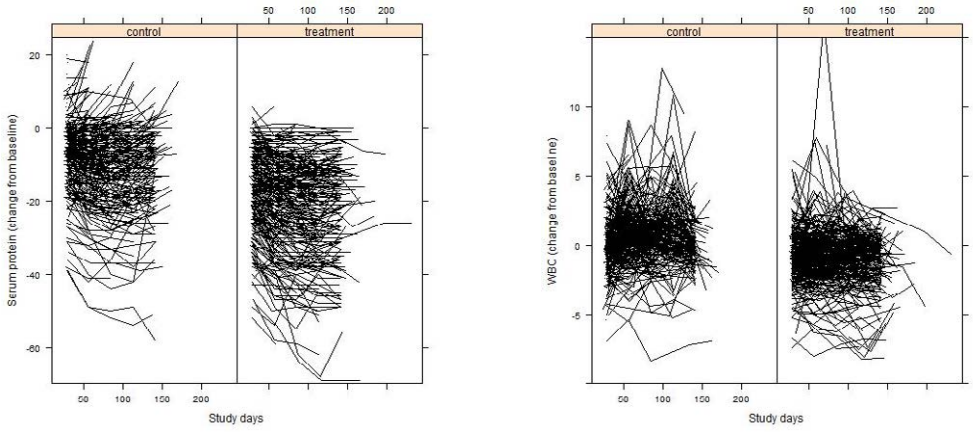

Figure 2: Change from Baseline of Serum Protein Level (left) and WBC (right) by Treatment Groups

The data for the change of serum protein level was analyzed using JM package in R and the result shows an association between longitudinal process and hazard rate of the event process $\alpha = 0.0236$, which indicates the larger reduction of serum

protein level will reduce more of the hazard rate. The data was also analyzed using another package in R, i.e. joineR. The results are similar but not identical and are not shown here.

```
Joint Model Summary:

Longitudinal Process: Linear mixed-effects model
Event Process: Relative risk model with piecewise-constant
               baseline risk function

Longitudinal Process
               Value Std.Err  z-value p-value
(Intercept) -10.0584  0.3024 -33.2590 <0.0001
day           0.0038  0.0031   1.2074  0.2273
day:trtgrp   -0.0346  0.0031 -11.0981 <0.0001

Event Process
            Value Std.Err  z-value p-value
trtgrp    -0.8966  0.1263  -7.0986 <0.0001
Assoct     0.0236  0.0044   5.3997 <0.0001
```

Similarly, the data for the change of WBC was analyzed and the result shows an association between longitudinal process and hazard rate of event process $\alpha = -0.0285$. Which indicates too much of the WBC reduction will actually increase the hazard rate. That should not be surprising as low level of WBC will increase the chance of infection and that can cause other complications.

```
Joint Model Summary:

Longitudinal Process: Linear mixed-effects model
Event Process: Relative risk model with piecewise-constant
               baseline risk function

Longitudinal Process
              Value Std.Err z-value p-value
(Intercept) -0.2173  0.0986 -2.2039  0.0275
day          0.0053  0.0012  4.4575 <0.0001
day:trtgrp  -0.0098  0.0012 -8.4021 <0.0001

Event Process
            Value Std.Err  z-value p-value
trtgrp    -1.2105  0.1336  -9.0624 <0.0001
Assoct    -0.0285  0.0233  -1.2227  0.2214
```

5 Combining the Associations

To combine the associations, we bootstrapped 500 samples and estimated the 500 pairs of association. A covariance matrix between associations of serum protein level and WBC was estimated. The weights to combine the associations can be estimated using the following equation

$$\beta = \frac{1 - \rho(\sigma_1/\sigma_2)}{1 - 2\rho(\sigma_1/\sigma_2) + (\sigma_1/\sigma_2)^2},$$

and the BLUE of the combined association can be estimated by

$$\mathcal{A} = (1 - \beta)\mathcal{A}_1 + \beta\mathcal{A}_2,$$

where $\mathcal{A}_1$ and $\mathcal{A}_2$ are the associations of serum protein and WBC, respectively.

The combined effect of the change from baseline of serum protein level and WBC on the hazard of PFS can also be visualized with the following graph (figure 3). The combination of changes below the cyan-colored plate will reduce the hazard of PFS, and all other combinations of changes above the plate will increase the hazard.

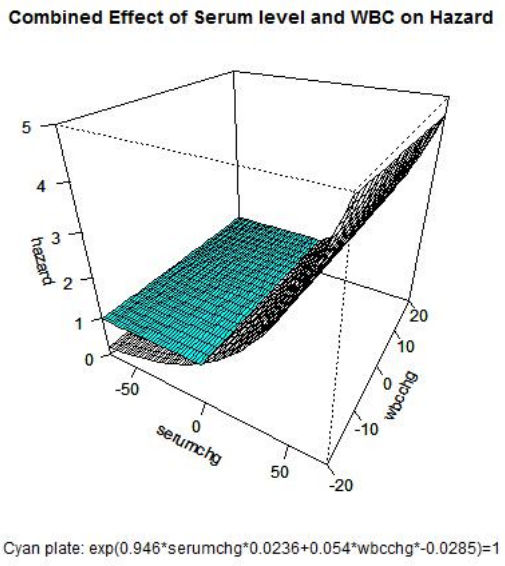

Figure 3: Joint effect of change from baseline of serum protein and WBC on the hazard of PFS

6 Summary

Clinical trials collect huge amount of data on efficacy and safety. To better understand the overall treatment effects, one needs to combine these data longitudinally as much as possible and to analyze efficacy and safety data, pre- and

post-treatment data together, as they quite often interact with each other. Joint model of various data types collected in clinical studies has been well-established in both theory and practices. We propose a general approach to combine the effect of two longitudinal data series and to estimate the joint effect on the hazard of the event process. This can easily be extended to more longitudinal data series and further work in under research. In addition, the results obtained using the two methods mentioned in this article are not always consistent. Whether this is data dependent or has intrinsic difference is unknown and more research will be conducted to compare these methods.

References

[1] Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data, Biostatistics, v1(4), p465-480.

[2] Therneau, T. and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

[3] R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[4] Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error, Biometrics, v53(1), p330-339.