PharmaSUG 2018 - Paper EP-02

Application of Survival Analysis in Multiple Events Using SAS

Jane Lu, AstraZeneca Pharmaceuticals, Gaithersburg, MD
David Shen, Independent Statistical Consultant

ABSTRACT

Cox proportional hazard model with time to first event as the outcome variable is commonly used to estimate the survival probability after adjusting for both baseline hazard and explanatory variables. This model counts subjects at risk at the time of first event. After that subjects are no longer considered to be at risk. There is a growing interest in analyzing recurrent or repeated events data in clinical studies. It has been proposed to extend survival models based on Cox proportional hazard approach to handle recurrent multiple events data. This paper introduces several different models in analyzing multiple events, compares and evaluates these models, and discusses their pros and cons. SAS codes are also provided for implementing these models.

INTRODUCTION

In clinical trials, Kaplan-Meier method is often used to estimate the survival function non-parametrically from observed (censored and uncensored) survival times. Cox proportional hazard model, the extensively used model, generates reliable estimates of survival times and hazard ratios associated with time-to-event data. As a semi-parametric model, it does not have any constraints on distribution assumption, which makes it an attractive alternative to parametric models.

Many applications involve repeated events, where a subject may experience multiple events over a trial duration or lifetime. However, the proportional hazard model can't handle recurrent multiple events because survival time in the standard model terminates at the time of the event. Only the time to first event is used and recurrent events are disregarded.

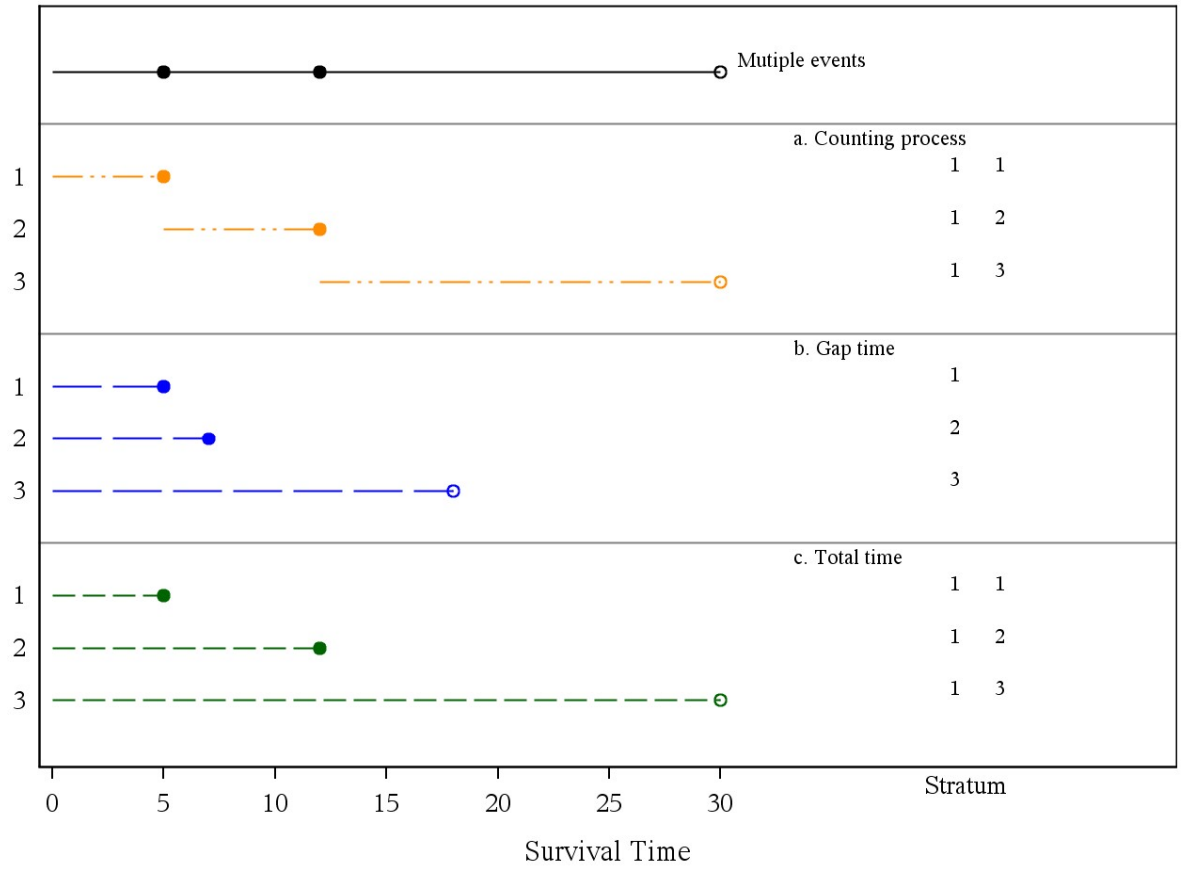
1. MODEL COMPONENTS

There are four components to a Cox-based multiple event model. These concepts are illustrated with the following example of a subject experiencing the events at day 5 and 12 separately, then censored at day 30, in a 30-day trial.

1.1 RISK INTERVALS

Risk intervals define the time when a subject is at risk of having an event along a given time scale.

Figure 1 Time to event data over the period of 30 days for one hypothetical subject



The scheme above illustrates three risk intervals: a) counting process, b) gap time and c) total time. Each time to an event or censoring is a separate risk interval.

1) Counting process uses the same time scale as total time, but recognizes that a subject may have a delayed entry or censored period before the subject becomes at risk for the event. In the counting process, subject is at risk for the first event, the second during (0, 5], (5, 12], respectively and then censored during (12, 30].
2) Gap time is the time from the prior event, that is, the clock restarts after each event. For example, with gap time, subject is at risk of the first event during (0, 5], and of the second event and censored during (0, 7] and (0, 18], respectively. In both gap time and counting process formulations, the subject is at risk for the same length of time. The risk interval for the first event is the same for all three risk intervals.
3) Total time is the time from a chosen point, usually the time from the start of treatment. With total time, subject is at risk for the first, second event and censored during (0, 5], (0, 12] and (0, 30], respectively.

The risk interval determines whether a model is either conditional or marginal. Counting process and gap time are conditional since a subject cannot be at risk until the end of the previous event, that is, the subject is at risk conditioned on previous events. Total time is marginal since the subject is a risk from the start of treatment and does not depend on any previous events.

1.2 BASELINE HAZARD

There are two baseline hazard functions available for multiple event models.

- Common baseline hazard
- Event-specific baseline hazard

A model with a common baseline hazard has the same underlying hazard for all events. An event-specific baseline hazard is a stratified baseline hazard that allows the baseline hazard to be different for each k^{th} event. Stratifying by event is essentially fitting a separate model for each k^{th} event.

1.3. RISK SET

The k^{th} risk set contains the individuals who are at risk for the k^{th} event. There are three possible risk sets: unrestricted, restricted and semi-restricted. The risk set definition incorporates the choice of baseline hazard. The risk set at a given point in time depends on the individuals included in the set and when those individuals are at risk, that is, the risk interval.

When the risk set is unrestricted, all the subject's risk intervals may contribute to the risk set for any given event, regardless of the number of events experienced by each subject. An unrestricted risk set has a common baseline hazard function for all events.

With a restricted risk set, contributions to the k^{th} risk set is restricted to only include the k^{th} event risk intervals of those subjects who have experienced (k-1) events. For example, only subjects who have already had three events will be considered to be at risk for the fourth event. A restricted risk set has event-specific baseline hazards.

Semi-restricted risk sets have event-specific baseline hazards but allow subjects who have fewer than (k-1) events to be at risk for the k^{th} event through the creation of 'dummy' risk intervals. Thus a subject who has had none or one event can be considered at risk of a fourth event. However, a semi-restricted risk set does not allow information from the k^{th} event risk interval to contribute to the risk set for an earlier event. This risk set only applies to the counting process and total time with event-specific baseline hazards.

1.4. WITHIN-SUBJECT CORRELATION

Three approaches have been proposed for accounting for the within-subject correlation between events: conditional, marginal and random effects. The conditional approach assumes that the current event is unaffected by earlier events. This assumption can be relaxed by introducing time-dependent covariates in the model, such as the number of prior recurrences, which may capture the dependence structure among the recurrence times. The marginal approach assumes that the events within a subject are independent and estimates a robust variance using a 'sandwich' estimator. The unadjusted variance estimates are called naive estimates. The random effects approach, also called frailty models, introduces a random covariate into the model that induces dependence among the recurrent event times.

2. MODELS

Of the four key components, risk interval and risk set are pivotal for constructing a model. Baseline hazard is linked to risk set definition. That is, an unrestricted risk set is necessarily associated with a common baseline hazard and the semi-restricted and restricted risk sets are necessarily associated with event specific hazards. These determine the interpretation of the resulting model. The possible extended Cox regression models to handle multiple events and clustered data include: AG, PWP-CP, PWP-GT, LWA and WLW.

1) Andersen and Gill (AG): $\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{X_{ij}(t)'\beta\}$
2) Prentice Williams and Peterson total time (PWP-CP): $\lambda_{ij}(t) = Y_{ij}(t)\lambda_{0j}(t)\exp\{X_{ij}(t)'\beta_j\}$
3) Prentice Williams and Peterson gap time(PWP-GT): $\lambda_{ij}(t) = Y_{ij}(t)\lambda_{0j}(t - t_j{-}1)\exp\{X_{ij}(t)'\beta_j\}$
4) Lee, Wei and Amato (LWA): $\lambda_{ij}(t) = Y_{ij}(t)\lambda_0(t)\exp\{X_{ij}(t)'\beta_j\}$
5) Wei, Lin and Weissfeld (WLW): $\lambda_{ij}(t) = Y_{ij}(t)\lambda_{0j}(t)\exp\{X_{ij}(t)'\beta_j\}$

$\lambda_{ij}(t)$ represents the hazard function for the *j*th event of the *i*th subject at time t;
$Y_{ij}(t)$ be the indicator variable for the *j*th event of the *i*th subject at time t, 1 when the subject is at risk and under observation, 0 otherwise;
λ_0 represents the common baseline hazard for all events;
λ_{0j} represents the event-specific baseline hazard for the *j*th event at time t;
$X_{ij}(t)$ represents the *j*th covariate vector for the *i*th subject at time t;
β_j be the parameter vector for the *j*th event, which includes the treatment effect.

2.1 COUNTING PROCESS MODEL

The AG model defines the risk intervals using counting process in which each event is assumed to be independent and a subject contributes to the risk set for an event as long as the subject is under observation at the time of the event. The data for each subject with multiple events can be described as data for multiple subjects where each has delayed entry and is followed until the next event. This model ignores the order of the events, leaving each subject to be at risk for any event as long as he is still under observation at the time of the event. This implies that a subject could be at risk for a subsequent event without having experienced the prior events.

The AG model is used with an unrestricted risk set and hence assumes a common baseline hazard for all events.

2.2 CONDITIONAL MODEL

The second type of model is conditional model. It assumes that a subject can't be at risk for event no. 2 without having experienced event no. 1 previously. A variable of stratum is used to indicate event order number, so this type of the models has event-specific baseline hazards and risk set is restricted.

The PWP-CP model (conditional model A) uses counting process formulation even though it is called a 'total time' model. The PWP-CP model is a stratified AG model with each time interval starting at the time of the previous event. The big difference between this model and the counting process model is that the stratum variable is used to keep track of the event number; thus, ensuring that it is not possible to be at risk for subsequent events without having experienced the previous events. In this model, the time interval of a subsequent event starts at the end of the time interval for the previous event. PWP-CP model is very useful for modeling the full-time course of the recurrent event process.

The PWP-GT model (conditional model B) differs from PWP-CP in that it uses gap time instead of counting process. In this model, each time interval starts at time zero and ends until the next event. As a result, the risk sets for these conditional models are completely different and the questions they answer are also very different. This model is very useful for modeling the time between each of the recurring events rather than the whole-time course of the recurrent event process.

Since the PWP models have event-specific baseline hazards, we can have either an overall estimate or event-specific estimates for each covariate.

Originally, the AG and PWP models do not adjust for the correlation within subjects, except by including covariates in the regression equation.

More recently, researchers have introduced the robust variance 'sandwich' estimator in an attempt to adjust for the within-subject correlation.

2.3 MARGINAL MODEL

In the marginal model, each event is considered as a separate process. The time for each event starts at the beginning of follow up time for each subject. Each subject is considered to be at risk for all events, regardless of how many events each subject actually experiences. All subjects in the study contribute follow up time to all possible recurrent events. Thus, the marginal model treats each event separately and

models all the available data for the specific event.

The LWA model employs total time and assumes a common baseline hazard with an unrestricted risk set. The LWA model allows a subject to be at risk for several events simultaneously. A subject with n risk intervals can contribute to a risk set n times. This is unique with this model, none of the other models allows this. The LWA model accounts for the within-subject correlation by adjusting the variance via the 'sandwich' estimator.

The WLW is an event-specific LWA model with a semi-restricted risk set. Since there is a maximum of three events, each subject is potentially at risk for three events, due to the semi-restricted risk set. Therefore, 'dummy' records are created so that there are three records for each subject. The estimates of WLW, like PWP, can be either event-specific or overall. However, the overall estimate proposed by WLW is defined differently; it is the weighted average of the event-specific estimates, $\beta_1,\ldots, \beta_K$. The WLW model also requires limitation of the data to a maximum number of events if event-specific estimates become unreliable.

TABLE 1. Summary of survival analysis models for multiple events

Model	Time interval	Censor	Stratum	Basic SAS codes

Counting process model
AG
	(0, 5]
	(5, 12]
	(12, 30]
		0
		0
		1
			1
			1
			1
				`proc phreg data=adef covm covs(aggregate);`
				`model (Tstart,Tstop)*Censor(1)= Trt01pn;`
				`id usubjid;`
				`run;`

Conditional model A
PWP-CP
	(0, 5]
	(5, 12]
	(12, 30]
		0
		0
		1
			1
			2
			3
				`proc phreg data=adef covm covs(aggregate);`
				`model (Tstart,Tstop)*Censor(1)= Trt01pn;`
				`strata stratum;`
				`id usubjid;`
				`run;`

Conditional model B
PWP-GT
	(0, 5]
	(0, 7]
	(0, 18]
		0
		0
		1
			1
			2
			3
				`proc phreg data=adef covm covs(aggregate);`
				`model GapTime *Censor(1)= Trt01pn;`
				`strata stratum;`
				`id usubjid;`
				`run;`

Marginal model
LWA
	(0, 5]
	(0, 12]
	(0, 30]
		0
		0
		1
			1
			1
			1
				`proc phreg data=adef covm covs(aggregate);`
				`model (Tstart,Tstop)*Censor(1)= Trt01pn;`
				`id usubjid;`
				`run;`

Marginal model
WLW
	(0, 5]
	(0, 12]
	(0, 30]
		0
		0
		1
			1
			2
			3
				`proc phreg data=adef covm covs(aggregate);`
				`model (Tstart,Tstop)*Censor(1)= Trt01pn;`
				`strata stratum;`
				`id usubjid;`
				`run;`

We can see that AG and LWA are unstratified models, while PWP-GT, PWP-CP and WLW are stratified

models

3. EXAMPLES

Analyses are conducted to investigate any biases and evaluate the inferences of the permissible models for recurrent event data with and without within-subject correlation.

Consider two different treatment scenarios:

1) treatment is constantly effective
2) treatment is effective only for the first event

When treatment is effective for the first event, the total time model displays a ‘carry-over' effect where treatment effect diminishes with each consecutive event.

If treatment is effective for one event, the total time to an event for a treated subject will always be longer compared to a placebo subject. However, as the time on treatment increases, the relative difference between treated and placebo subjects will get smaller, and so a gradual decrease in treatment effect is observed.

3.1. CONSTANT TREATMENT EFFECT

3.1.1. Independent events

First consider when treatment is constant and there is no induced within-subject correlation

1) The event-specific treatment estimates for PWP-GT and PWP-CP models have very similar results with negligible bias. The total time models overestimate the treatment effect after the first event, with the over-estimation becoming larger with each consecutive event in the WLW model. WLW concludes that for the 4th event treatment is over twice as effective than it is in reality.

   For all event-specific estimates the robust standard errors are similar to the naive standard errors. Event-specific standard errors between models are also similar. Hence, the empirical coverage probabilities of the confidence intervals are close to the nominal 95 percent level for all event-specific estimates for PWP-GT and PWP-CP models but is lower for the total time models, especially WLW.

2) The common and weighted average estimates in the event-specific models are in general very similar. These overall treatment estimates reflect the ‘average' of the event-specific estimates. Only the WLW model had a statistically significant difference between the two overall treatment effects, with the weighted average being less biased and having a better empirical coverage of the nominal confidence interval level. The robust standard errors are similar to the naive standard errors for the common treatment estimate in the PWP-GT and PWP-CP models, but are inflated for the total time models.

3) For the unrestricted models, the common treatment estimate for AG are similar to their corresponding restricted models, PWP-GT and PWP-CP, and have minimal bias. LWA, on the other hand, underestimates the treatment effect which is opposite to the other total time models, WLW. The robust standard errors are the same as the naive standard errors in all unrestricted models.

3.1.2. Within-subject correlation

Within-subject correlation decreases treatment estimates as compared to the corresponding model estimates when the events are independent. As within-subject correlation increases, the smaller the treatment estimates become. All models are affected by within-subject correlation. The robust standard errors are similar to the naive standard errors for the event-specific estimates. There is a small increase in the robust standard error for the common and weighted average estimates for PWP-GT and the

unrestricted models, AG and LWA, but not for PWP-CP. However, the inflated standard error is insufficient to compensate for the bias in the estimated treatment effect.

3.2. TREATMENT EFFECTIVE FOR FIRST EVENT ONLY

3.2.1. Independent events

Consider when events are independent. PWP-GT and PWP-CP show that treatment is effective for the first event only, with minimal bias and empirical coverage close to the nominal level.
The WLW overestimates treatment effect after the first event and hence empirical coverage is lower than the nominal level.

When treatment effect is not constant, the unrestricted models make little sense and do not reflect the expected ‘average' treatment effect as estimated by the weighted average and common estimates of the restricted models.

Even for the event-specific models, using an overall estimate of treatment effect when treatment effect is not constant between events probably has little value and should be interpreted cautiously. If it is used, it should be stated clearly that treatment effect is not constant.

3.2.2. Within-subject correlation

Inducing within-subject correlation has the same effect on estimates as before when treatment is constant; it underestimates the treatment effect compared to the corresponding model when events are independent.

3.3. COMMENTS ON THE RESULTS

Recurrent event data have two closely linked features: the subject can only be at risk for one event at a time; and the events are ordered, for example, the first event occurs before the second event, where the hazards may change after each event. The choice of the model characteristics should be determined by this data structure and research topic of interest. How the model characteristics, risk interval and risk set are defined determines the results and their interpretation.

3.3.1. Risk Interval Definition

Total time has a ‘carry-over' effect. Total time within a subject tends to be highly correlated. For instance, the total time of the second risk interval includes the first interval; the third risk interval contains the first and second intervals, and so on. An analysis based on total time which estimates a large treatment effect for the first event will carry over this effect for subsequent events as seen with the simulated data where treatment is effective for the first event only, even when an analysis of gap times suggests that the treatment is no longer effective. When treatment effect is constant, it is expected that the event-specific estimates in a total time model are constant.
Counting process estimates is extremely close to gap time estimates. Subjects for these models are at risk for the same length of time but on a different time scale. The counting process estimates are not close to the total time estimates, which share the same time scale but differ in the length of time at risk. Hence, the length of time at risk may influence the results much more than the time scale. However, using gap time has a different hazard ratio interpretation compared to total time and counting process estimates because of the different time scale - for gap time the hazards are since the last event, and for total time or counting process the hazards are since start of treatment.

3.3.2. Risk Set Definition

Semi-restricted risk sets have a carry-over effect. This is clear from comparing the results of WLW. For example, if treatment is effective then treated subjects will have fewer events in the same period. A semi-restricted risk set includes all subjects in each event-stratum, and so with each consecutive event the

number of treated subjects with a censored observation increases.

These censored observations are compared to those untreated subjects who are experiencing an event, and so exaggerate the treatment effect in the later strata.

A restricted risk set is essential for recurrent event data if the hazards are expected to change after each event. For our results, it was important to stratify when the treatment effect was not constant, otherwise the model estimate did not reflect the 'average' treatment effect. Event-specific models also have the advantage of observing how effects change with each subsequent event.

An unrestricted risk set is suitable when the baseline hazard does not change with each event. However, this is only appropriate for the AG models and not for LWA. As a result, the LWA was the only model to underestimate the treatment effect because it was the only model that allowed a subject to be at risk several times for the same event.

3.3.3. Robust Variance

A deflated robust standard error means there is more variation within subjects than between subjects, while an inflated robust standard error means there is less variation within subjects than between subjects.

From the simulated results, the robust standard errors of the weighted average and common estimates in the total time models are inflated regardless whether events are independent within subjects or not. For all the other models, except PWP-CP, the robust standard errors of the weighted average and common estimates are the same as the naive standard errors when within-subject events are independent, and become inflated when events are not independent. The naive and robust standard errors for the weighted average and common estimates for PWP-CP appear to be the same regardless of the within-subject correlation.

3.3.4. Software

The models discussed can be implemented in various software packages. For example, SAS version 9.2 or later. The SAS 6.10 STAT Enhancements and Changes manual clearly explains how to use PROC PHREG to fit the AG, WLW and PWP models. The variance 'sandwich' estimator is a little clumsy, requiring PROC IML.

4. DISCUSSION AND CONCLUSION

We recommend using the PWP-GT model to analyze recurrent event data when within-subject events are independent. Each model answers a slightly different research question. The PWP-GT model would be used to determine whether the treatment is effective for the k^{th} event from the time since the previous event. The TT-R model can be used to assess whether treatment is effective for the k^{th} event since the start of treatment. The AG models are adequate if a common baseline hazard can be assumed, but they lack the details and versatility of the event-specific models.
Applying a robust variance may not be adequate when there is within-subject correlation.
Further work is required to investigate when and why the robust variance estimate fails to adjust for within-subject correlation. Perhaps other methods, such as frailty models, should be used instead.

The WLW and LWA are not appropriate for recurrent event data. WLW has a semi-restricted risk set that allows subjects to be at risk of the 4th event even if the subjects have only had one event, which leads to overestimation of treatment effect. LWA allows subjects to be at risk several times for the same event.

However, these models are suitable in other situations. The WLW is most appropriate to data when there are different types of events for the same person, where the baseline hazard is potentially different for each type, such as multi-type event data, for example, tumors at several different sites in the body. The

LWA model is suitable for clustered data such as siblings or pairs of eyes, where it can be assumed that the baseline hazard is the same, and the beginning of risk of the event is the same within the cluster.

The LWA model is not appropriate for recurrent event data because it allows a subject to be at risk several times for the same event.

The WLW model overestimates treatment effect and is not recommended.
We conclude that PWP-GT is useful model for analyzing recurrent event data, providing answers to slightly different research questions.

Further, applying a robust variance to any of these models does not adequately account for within-subject correlation.

CONTACT INFORMATION

Jane Lu

Janelu6@hotmail.com
AstraZeneca Pharmaceuticals
Gaithersburg, MD