Recurrent Event Analyses Illustrated in the Pivotal Exacerbation Study SPARK in the Respiratory Area

Hua Li[1], Paul Gallo[1], Richard Cook[2]

[1]Novartis Pharmaceuticals Corporation, 1 Health Plaza, NJ 07936
[2]University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

Abstract

For many clinical trials in the respiratory area, the primary response is based on potentially recurrent exacerbations observed during a treatment period. Conventionally, logrank tests and Cox proportional hazards models have been used to compare treatments with regard to the time to the first exacerbation and a parametric time-homogeneous negative binomial model has been used to estimate the exacerbation rate. These analyses do not make use of potentially important information, namely the times of event occurrence, which can improve understanding of the exacerbation process and enhance understanding of the treatment benefits; one can also base analyses on methods which are robust to model misspecification. In this paper, we review a number of approaches reflecting recent advances in recurrent event methodology, including the marginal Cox model, rate-based models and the semiparametric negative binomial model. We demonstrate the application of these methods to the pivotal SPARK study conducted for approval of a drug for COPD. We discuss the strengths and limitations of the competing methods and the interpretation of the findings from the various analyses before making recommendations on approaches for the design and analysis of future trials in COPD.

Key Words: Recurrent event data analysis, marginal Cox model, rate-based model, semiparametric negative binomial model

1. Introduction

In many therapeutic areas and disease conditions, the outcome of interest that we hope to alleviate using a novel therapy may occur more than once within individual patients. An important example in the respiratory area involves exacerbation studies in treatments for COPD (Chronic Obstructive Pulmonary Disease) or asthma, where the main outcome is an exacerbation event which may occur multiple times during the study for individual patients. Conventionally, logrank tests or Cox proportional hazards (PH) models have been used to compare treatments with regard to the time to the first exacerbation event and a parametric negative binomial model has been used with regard to the exacerbation rate. However, these analyses do not make use of potentially important information, namely the total number and/or times of occurrence of repeat exacerbations, which might potentially be used to improve the power of the analysis and the ability to most accurately characterize treatment benefits and be more robust to model misspecification. More sophisticated analysis methods have been developed to analyze recurrent event data, and these are slowly gaining increased credibility with regulators. In this paper, we describe and review a number of approaches reflecting recent advances in recurrent event methodology, including methods based on marginal Cox models (Wei, Lin, and Weissfeld, 1989), models based on rate or mean functions (Anderson and Gill, 1982; Lawless and Nadeau, 1995; Lin et al, 2000) and the semiparametric negative binomial model (Therneau and Grambsch, 2000). We apply these methods to a pivotal study SPARK conducted for approval of a new drug for COPD and discuss the advantages and limitations of the various methods.

Ultibro® Breezhaler® (QVA) is a fixed-dose combination of indacaterol maleate (a long acting beta-2 agonist) and glycopyrronium bromide (NVA, a long acting muscarinic antagonist) and was developed for the once-daily treatment of COPD. The SPARK study was a 64-week, multi-center, randomized, double-blind parallel-group, active controlled study to evaluate the effect of QVA vs NVA and open-label tiotropium (Tio, 18 μg o.d.) on COPD exacerbations in patients with severe to very severe COPD. The study involved three arms, QVA, NVA, and Tio with sample sizes of around 730 patients per arm. The primary analysis variable of the study was the rate (time-adjusted numbers) of adjudicated moderate or severe COPD exacerbations during the treatment period (period between the first day of the study drug administration to the last day of the study drug administration). Out of 2205 patients in the modified full analysis set, 1247 had exacerbations and more than 50% of patients experiencing exacerbation had multiple exacerbations (up to 11 events); see Figure 1.

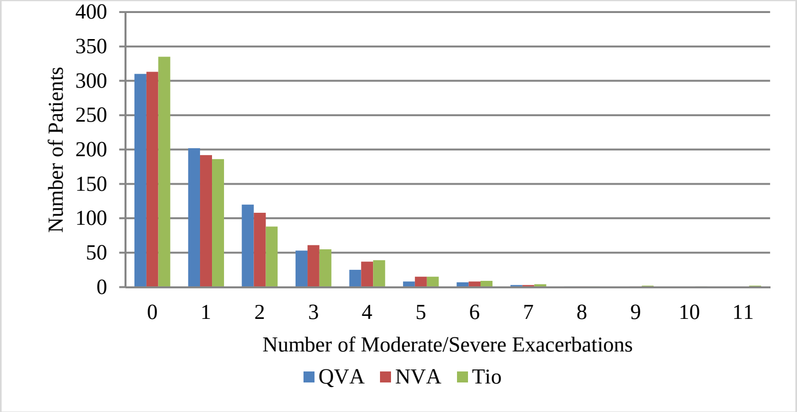

Figure 1: Histogram of the number of exacerbations per patient by treatment arm

The administrative censoring time was after last study drug administration if the patient remained in the study without early withdrawal. Even though there were some early dropouts, the overall censoring rate was very low (approximately 4.2%, 6.4% and 4.3% in groups QVA, NVA and Tio, respectively), as shown by the Kaplan-Meier plot in Figure 2.

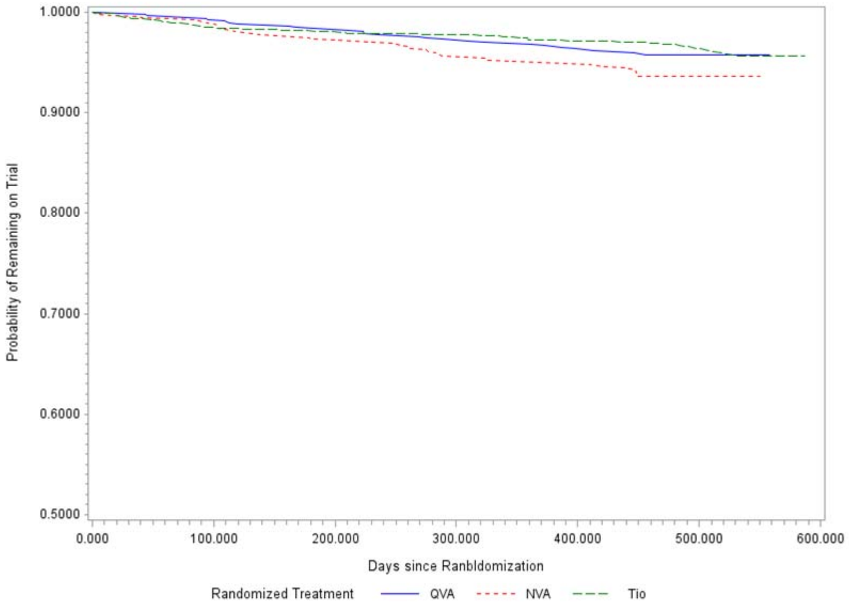

Figure 2: Kaplan-Meier estimates of the time on study

We plot the log of the Nelson-Aalen estimates of the cumulative mean functions (CMF) for each arm of the trial to help assess the suitability of the proportional means model; these estimates should be roughly parallel if the proportionality assumption is correct for the rate or mean functions. Figure 3 does not suggest any violations of this assumption.

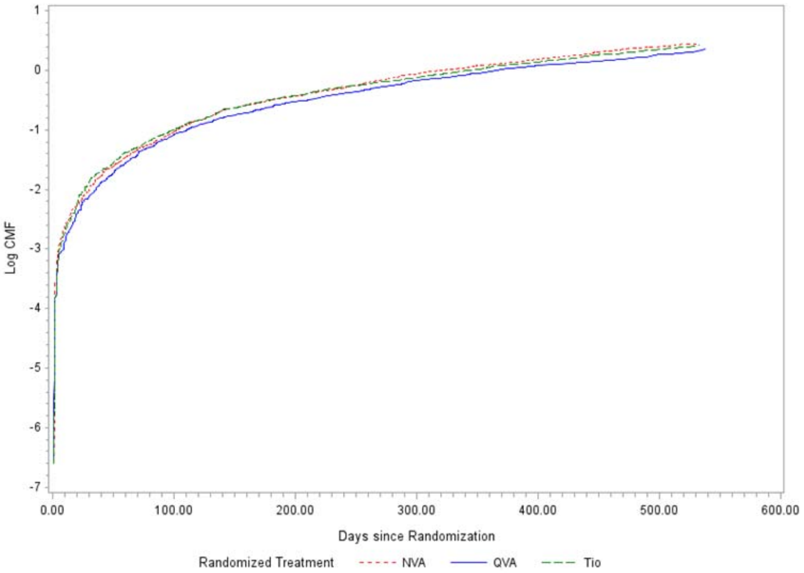

Figure 3: Log of cumulative mean functions for exacerbations versus time

When mortality rates are high, a model that addresses both the recurrent event and the mortality processes might be needed. However in this study, the death rate is very low at about 3% during the treatment period. So death is not formally dealt with as a terminal event. This paper focuses on the application of the recurrent event methods to the exacerbation data.

2. Recurrent Event Models

2.1 Notation

Following the general notation used in Cook and Lawless (2007), let $t_{i1} < t_{i2} < \ldots$ denote a sequence of recurrence times for subject *i*, $N_{ik}(t) = I(t_{ik} \le t)$ indicate that the k^{th} event occurred over (0, t], and $N_i(t) = \sum_{k=1}^{\infty} I(t_{ik} \le t)$ count the total number of events over (0, t], for individual *i* in a sample of size n, *i*=1,…, n. Let $H_i(t) = \{N_i(s), 0 < s < t, x_i\}$ denote the history of the recurrent event process up to time t and $H_{ik}(t) = \{N_{ik}(s), 0 < s < t, x_i\}$ the history of the event-specific process up to time t.

The Cox PH model for the 1st exacerbation is

$$\lim_{\Delta t \downarrow 0} \frac{P\{\Delta N_{i1}(t) = 1 \mid H_{i1}(t)\}}{\Delta t} = I(t \le t_{i1}) * h_1(t \mid x_i),$$

where we specify $h_1(t \mid x) = h_{01}(t)\exp(x\beta_1)$ and estimate the hazard ratio $\exp(\beta_1)$.

2.2 Methods for Recurrent Event Data

In exacerbation studies, traditionally two types of endpoints are of clinical interest. One is time to event endpoint, for example the time to the first exacerbation, which provides information on the delay of the event via the hazard ratio. For the time to event endpoint, when multiple events are to be considered we focused on the Wei, Lin, and Weissfeld (1989) marginal Cox PH model; we discuss this further in the next section. The other approach is to consider the recurrent event endpoints which are usually summarized in terms of an annualized rate and its reduction in the form of the rate ratio. More sophisticated semiparametric analyses are based on the Anderson-Gill method and its robust versions (Lawless and Nadeau, 1995; Lin et al., 2000) or the semiparametric negative binomial model.

The gap time analysis is not discussed here, as this is generally not recommended for randomized trials where causal inference is a priority. For instance, the Prentice et al. (1981) gap time model is a stratified Cox regression model based on the prior number of events. Typically in a randomized clinical trial, after experiencing the k^{th} (*k*=1,…,K-1) exacerbation the treatment groups may not necessarily be balanced anymore with regard to baseline characteristics; therefore the comparison among treatment arms may be questionable. In addition, the treatment effects may confound with the prior event history, which also makes the interpretation of treatment effect more difficult.

2.2.1 The robust Wei, Lin and Weissfeld approach

For time to event endpoints, traditionally the Cox PH model and logrank test have been used. Wei, Lin and Weissfeld (1989) extended this approach by simultaneously modelling the marginal distribution of the time to each of several different clinical events with a Cox PH model. In the present setting these events represent the successive exacerbations up to a final exacerbation to be modelled.

The hazard for the k^{th} (*k*=1…K) exacerbation is

$$\lim_{\Delta t \downarrow 0} \frac{P\{\Delta N_{ik}(t) = 1 \mid H_{ik}(t)\}}{\Delta t} = I(t \le t_{ik}) * h_k(t \mid x_i),$$

and $h_k(t \mid x) = h_{0k}(t)\exp(x\beta_k)$.

The Wei, Lin and Weissfeld (WLW) approach does not impose a particular structure of dependence among distinct exacerbation times on each subject, but uses the robust variance estimation to take care of the within-patient dependence in the event times. Asymptotically the resulting estimators of β_k (k=1…K) are jointly normally distributed. This approach offers the opportunity to study the treatment effects over time and treatment effects averaged over the events. Implementation through a single model is available in the SAS PROC PHREG procedure via the STRATA statement. In this formulation, subjects are considered "at risk" for their k^{th} event irrespective of whether they have experienced their $(k-1)^{st}$ event; this neglects the natural ordering of the recurrence times, and has been criticized by reviewers and practitioners (Rejoinders, 1992; Cook and Lawless, 1997; Tuli et al., 2000; Metcalfe and Thompson, 2007).

To perform this analysis, with K=4, each patient should have 4 records in the data file of the counting process format (Therneau and Grambsch, 2000). For a patient with fewer than 4 events, dummy records need to be created for the remaining events. For example, for a patient with only 2 events as shown in Table 1, the 3rd and 4th dummy records are created as in Table 2 so that this patient will be in the "risk set" for these events.

Table 1: Patient with two events

Patient ID	Exacerbation number	Start	Stop	Status
0003	1	1	83	1
0003	2	98	502	1

Table 2: Augmented records created for the third and fourth event times

Patient ID	Exacerbation number	Start	Stop	Status
0003	1	1	83	1
0003	2	98	502	1
0003	3	509	533	0
0003	4	533	533	0

In this study the value of K was set to 4, since only 7% of the patients who experienced exacerbations had 5+ events.

2.2.2 Anderson-Gill model

Anderson and Gill (1982) introduced a counting process model based on the Poisson distribution where the intensity function $\lambda(t \mid H_i(t))$ has the same Cox-type form,

$$\lambda(t \mid H_i(t)) = \lim_{\Delta t \downarrow 0} \frac{P\{\Delta N_i(t) = 1 \mid H_i(t)\}}{\Delta t} = \lambda_0(t)\exp(x_i\beta),$$

where $\lambda_0(t)$ is the baseline intensity function.

Lawless and Nadeau (1995) and Lin et al (2000) extended the utility of this model by developing robust variance estimates to adjust for extra-Poisson variation, or equivalently the dependencies among multiple events within patients. These are also called rate and mean function models, and are most suitable when dealing with fixed or external time-varying covariates.

As discussed by Cook and Lawless (2007, Chap. 3 and Sec. 8.4), the use of rate and mean functions and the Anderson-Gill (AG) model offers the most broadly appealing and applicable approach for the comparison of treatment arms in randomized trials, provided censoring is independent of event occurrence and that there are no terminating events.

2.2.3 Semiparametric negative binomial model

For the analysis of exacerbation rates, traditionally a parametric negative binomial analysis has been performed based on event counts. The total number of events over the treatment period is assumed to follow a negative binomial distribution and the log of the treatment duration (exposure time) is included in the model as the offset along with other covariates. These models fully specify but ignore the information regarding when the recurrent events occur.

A semiparametric negative binomial model can also be used to address the heterogeneity among patients while the events within the same patient follow a Poisson distribution. In this case one can define a “subject-specific” intensity function as

$$\lambda(t \mid H_i(t), \gamma_i) = \gamma_i \lambda_i(t),$$

where γ_i is a random effect with $E(\gamma_i) = 1$ and $Var(\gamma_i) = \phi$ giving $E(N_i(t) \mid x_i) = \mu_i(t)$ and $Var(N_i(t) \mid x_i) = \mu_i(t) + \mu_i(t)^2 \phi$ where $\mu_i(t) = \int_0^t \lambda_i(u) du$ (Lawless, 1987; Therneau and Grambsch, 2000). The model is called semiparametric since $\lambda_0(t)$ is not assumed to have any parametric form. If $\phi = 0$, then this model reduces to a Poisson process, but generally it is a more flexible model. If $\gamma_i \sim Gamma(1, \phi)$, then this becomes a Gamma-Poisson mixture, i.e., a negative binomial process. In this case maximum likelihood estimates can be obtained even for the semiparametric model via SAS 9.3 or above (with STAT 13.2) using PROC PHREG with the RANDOM statement with DIST=gamma.

3. Assessment of Independent Censoring

The marginal analysis based on the Anderson-Gill model with robust variance estimates is valid under independent censoring as mentioned above. When there is a concern about dependent censoring (i.e., when individuals experiencing high rates of events are at higher risk of withdrawal), biases may arise in the AG analysis and analyses based on fully specified distributional assumptions, such as parametric negative binomial and semiparametric negative binomial models, may be preferred.

Here we focus on two types of dependent censoring: covariate-dependent censoring and event-dependent censoring.

3.1 Covariate-dependent censoring

Covariate-dependent censoring arises if there are covariates that are associated with both the recurrent event process and the censoring process, but are not controlled for in the recurrent event analysis. In that case we have to model the covariate effect on the censoring process and weight the AG estimating equations by the inverse of the probability of remaining on study conditional on the covariate that is associated with both the recurrent event process and the censoring process.

If we are just looking at a simple treatment comparison and only adjust for treatment in the rate model, then anything else associated with the risk of recurrent events that is also associated with the risk of censoring would qualify as a covariate that induces covariate-dependent censoring. For example, baseline FEV_1 is associated with risk of exacerbations as well as risk of censoring, since sicker patients might be more likely to be withdrawn from the study. If we only control for the randomized treatment in the rate function analysis then the omission of baseline FEV_1 induces dependent censoring. If we include baseline FEV_1 in the rate function model then it is no longer omitted and we are fine. But if we omit it, we would model the hazard for censoring given baseline FEV_1 and compute the censoring weight conditional on baseline FEV_1. This, if it is modelled correctly, deals with the dependent censoring and will ensure consistent estimation of the mean functions and the treatment effect.

3.2 Event-dependent censoring

For event-dependent censoring, the “covariate” mentioned above is time-dependent and is in fact the recurrent event process itself. If events predict more events, and events predict censoring, then we have this same issue. The models for assessing event-dependent censoring are typically fitted by creating time-dependent covariates based on an evolving collection of information on a response along with fixed baseline data.

When accessing the covariate-dependent censoring, a Cox regression model is utilized where the dependent variable is the censoring time with status event=1 indicating censoring and 0 otherwise. The model includes fixed covariates which are considered to be associated with the exacerbations: treatment, smoking history, COPD exacerbation history, total daily symptom score at baseline, inhaled corticosteroids, and FEV_1 before inhalation. Out of the 6 covariates, three covariates - COPD exacerbation history, total daily symptom score at baseline, and inhaled corticosteroids - turn out to have a significant association with the time of patient withdrawal ($p < 0.0001$), which suggests that there is some degree of covariate-dependent censoring. However since these covariates are all included in the pre-specified response model and no additional variables are identified to have potential association with the exacerbation occurring or the censoring process, the model is considered fine.

When accessing the event-dependent censoring, two models have been performed. One includes the number of events as a continuous variable; and the other includes indicator variables $z_i(i=1...5)$ with $z_i(i=1...4)$ indicating the i^{th} event and z_5 the 5+ events, i.e., treats the event numbers as discrete variables. Both models include the fixed covariates as mentioned above. Both models show that censoring time depends on the number of events or the indicator variables $z_i(i=1...5)$ with $p<0.0001$. However the overall censoring rate is so low that later analysis shows that there is a negligible impact on the estimates from the recurrent event analyses, as shown in Table 4.

4. Application to the SPARK COPD study

The results provided below are based on the moderate/severe exacerbations originating from the eCRF (electronic Case Report Form), not based on the adjudicated moderate/severe exacerbations dataset since data were not available for counting type of data analyses.

For the time-to-event analyses, the WLW model analyses the time to 1st, 2nd, 3rd, and 4th exacerbations instead of using the 1st exacerbation only. It confirms the result for the time to the 1st exacerbation using the regular Cox PH model, as shown in Table 3. Even though QVA and NVA show similar treatment effect with a hazard ratio of 0.94 for the time to the 1st exacerbation, for subsequent events, there is a clear trend that QVA delays the time to exacerbation more positively for later events, with the hazard ratio of QVA vs. NVA varying from 0.85 to 0.54 for the time to the 2nd exacerbation to the time to the 4th exacerbation. The overall treatment effect, averaged across 4 events, failed to demonstrate a significant effect.

Table 3: Time to first and first four event analyses via the WLW approach

	Hazard Ratio (p-value) of QVA vs. NVA				
	1st Exac.	2nd Exac.	3rd Exac.	4th Exac.	Global
Cox PH model	0.94(0.351)	NA	NA	NA	NA
WLW method	0.94(0.352)	0.85(0.136)	0.72(0.030)	0.54(0.005)	0.93(0.271)

For the exacerbation rate analyses, all three methods provide similar estimates for the rate ratio and confidence intervals as shown in Table 4. Semiparametric negative binomial model provides the smallest p-value with the narrowest CI among the three methods. The reasons for this could be many fold, for instance, some degree of dependent censoring, non-constant risk from patient to patient, the gap times being quite large for some exacerbations, or other unknown reasons. To determine if this result is driven by the data or there is systematic reason in the respiratory exacerbation data so that semiparametric negative binomial model consistently provides smaller standard deviation compared to the other two models, further work is needed. However the consistent estimates of the rate ratio from the three methods indicate that the event-dependent censoring noticed earlier is negligible, primarily due to the low censoring rate.

Table 4: Recurrent event analyses based on rate function model for QVA vs. NVA

	Rate Ratio (95% CI)	p-value
Parametric negative binomial model	0.88 (0.777, 0.997)	0.045
Semiparametric negative binomial model	0.87 (0.792, 0.961)	0.006
AG model with robust variance estimate	0.87 (0.771, 0.989)	0.033

In general, the semiparametric negative binomial model provides some protection for event-dependent censoring and does not require constant rate across patients, thus is recommended when event-dependent censoring is a concern. The AG approach does not require the counts to follow any specific underlying distribution as compared to the parametric negative binomial model, nor require any specification for the random effect distribution as compared to the semiparametric model, and therefore is more robust. Thus AG model is recommended when it is reasonable to assume independent censoring or if censoring rates are very low.

5. Concluding Remarks

The WLW approach needs pre-specification of the maximum number of events K (usually small) to be analyzed. In this study, the 5th and above recurrent events are therefore ignored in the WLW analysis. This truncation represents a loss of 7% of the recurrent event data, which means the treatment effect averaged across 4 events does not exploit the entire recurrent event process. We also note that the patients are considered "at risk" for their k^{th} event irrespective of whether they have experienced their $(k-1)^{th}$ event, which may raise concern regarding interpretation.

In clinical trials it is particularly important that models most accurately characterize treatment benefits, and that treatment effects are easily interpreted and understood. In settings involving very few events per subject it may be reasonable to focus simply on the time to the first event. However when events occur more frequently, as in the SPARK trial, the preferred approach is to utilize data from the full event processes and base analyses on the rate of exacerbations.

Estimates of the rate reduction via the AG model are easily interpreted and understood and robust variance estimates, which does not require any particular underlying distribution for the event rate, generally offer protection against extra-Poisson variation and other general departure from the Poisson model. It therefore is the most appealing and the simplest specification of treatment effects for recurrent events under the independent censoring assumption. The semiparametric negative binomial model provides some protection for event-dependent censoring and is recommended when this is a concern.

References

1. P.K. Anderson, R.D. Gill (1982), Cox's regression model for counting processes: a large sample study, The Annals of Statistics, Vol. 10, No. 4, 1100-1120
2. J. Boher, R.J. Cook (2006), Implication of model misspecification in robust tests for recurrent events, Lifetime Data Analysis 12:69-95
3. R.J. Cook, J.F. Lawless (1997). Discussion of paper by Wei and Glidden. Statistics in Medicine; 16: 841–3.
4. R.J. Cook, J. F. Lawless (2007), The statistical analysis of recurrent events. Springer
5. J.F. Lawless and C. Nadeau (1995), Some simple robust methods for the analysis of recurrent events, Technometrics, Vol. 37, No. 2, 158-168
6. D.Y. Lin, L.J. Wei, I. Yang, and Z. Ying (2000), Semiparametric regression for the mean and rate function of recurrent events, Journal of the Royal Statistical Society, Series B Vol. 62, Part 4, 711-730
7. C. Metcalfe, S.G. Thompson (2007) Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: should it be applied to a recurrent events outcome?, Statistical Methods in Medical Research; 16: 103–122
8. O. D. Rejoinders (1992). In Klein JP, Goel PK eds. Survival analysis: state of the art. Kluwer Academic Publishers.
9. T.M. Therneau, P.M. Grambsch (2000). Modeling survival data: extending the Cox model. Springer, New York.
10. S. Tuli, J. Drake, J. Lawless, M. Wigg, M. Lamberti-Pasculli (2000). Risk factors for repeated cerebrospinal shunt failures in pediatric patients with hydrocephalus. Journal of Neurosurgery; 92: 31–8.

11. L.J. Wei, D.Y. Lin, and L. Weissfeld (1989), Regression analysis of multivariate iIncomplete failure time data by modeling marginal distributions, Journal of the American Statistical Association, Vol. 84, No.408, 1065-1073