A Semi-Markov Model for Survival Data with Covariates

SHU-CHEN WU
Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109

Received 12 June 1981; revised 10 February 1982

ABSTRACT

Clinical trials are often concerned with the evaluation of two or more time-dependent stochastic events and their relationship. The information on covariates for individuals in the studies is valuable in assessing the survival function. This paper develops a multistate stochastic survival model which incorporates covariates. It is assumed that the underlying process follows a semi-Markov model. The proportional hazards techniques are applied to estimate the force of transition in the process. The maximum likelihood estimators are derived along with the survival function for competing risks problems. An application is given to analyzing the survival of patients in the Stanford Heart Transplant Program.

1. INTRODUCTION

In prospective studies and clinical trials, study subjects may make transitions among finitely many well-defined states. For example, in cancer clinical trials, possible states are improvement, partial response, complete response, and progression. The information available is a sequence of occupied states and the sojourn time in each of these states. Information about covariates is always available. We would expect a more comprehensive and instructive evaluation of the survival function from incorporating this multistate information with covariates into the model.

In this paper, we consider a stochastic model based on a semi-Markov process to describe the multistate, partially censored survival data. We also utilize Cox's [2] proportional hazards model of explanatory variables, or covariates, to estimate the force of transition of the semi-Markov process. With regard to multistate survival data, several studies about the implication of semi-Markov models have appeared in the literature (e.g., Weiss and Zelen [14]; Lagakos, Sommer, and Zelen [7]). Weiss and Zelen [14] proposed a semi-Markov model for clinical trials, but they did not take into consideration the right-censored observations. The nonparametric likelihood methods

52 Vanderbilt Ave., New York, NY 10017 0025-5564/82/06197+10$02.75

proposed by Lagakos et al. [7] provide estimates with several desired properties; however, these are not directly applicable to situations where covariate information is suitable for use. We propose here a stochastic model which permits an arbitrary number of transient and absorbing states and makes provision for censored data as well as covariate information. To accommodate censored data, we consider censoring as a cause of failure (see Prentice et al. [11]). For such a case, censoring may be referred to as loss of follow-up, or incomplete information due to the termination of the study. In other words, censoring is treated as one of the absorbing states. As a competing risk problem, this formulation would enable us to estimate the marginal distributions, or subsurvival functions (Peterson [10]). Peterson discussed the relationship between the overall survival and subsurvival functions for the case of two states: censoring and a single cause of failure. A similar argument extended to competing risk problems is proposed in Equation (7).

In Section 2, we discuss the probability model associated with a semi-Markov process for the multistate clinical trial. In Section 3, the maximum likelihood estimators of parameters and their asymptotic properties are derived. Section 4 analyzes the data from the Stanford Heart Transplant Program by applying our model to estimate the posttransplant survival function.

2. MODEL

Consider a semi-Markov model with a finite number of states which can be decomposed into two mutually exclusive subsets, A and Υ. The set of states A corresponds to absorbing states, and Υ to transient states. Set A includes the well-defined endpoint events, such as dying of the cause under study, dying of other diseases, withdrawing, or being censored by the termination of the study. The transient states include all remaining events.

Let J_l denote the state corresponding to the lth transition, with J_0 the initial state, and let the random variable T_l denote the duration of sojourn time between the $(l-1)$th and lth transitions. From the constructive definition of a semi-Markov process given by Pyke [12], the observed history of a process having m transitions to an endpoint event can be denoted as

$$H_m = \{J_0, T_1, J_1, \ldots, T_m, J_m\}, \tag{1}$$

where $J_l \in \Upsilon$ for $l < m$ and $J_m \in A$. In practice, H_m is the history of a patient who is observed up to the occurrence of one of the well-defined endpoint events in m transitions.

A semi-Markov process can be characterized by (a) the initial probability θ_{j_0}, (b) the transition probability p_{ij}, and (c) the right-continuous distribution

function $F_{ij}(t)$, defined respectively by

$$P\{J_0 = j_0\} = \theta_{j_0},$$
$$P\{J_l = j | J_0 = j_0, \ldots, J_{l-1} = i\} = P\{J_l = j | J_{l-1} = i\} = p_{ij},$$
$$P\{T_l \leqslant t | J_0 = j_0, \ldots, J_{l-1} = i, J_l = j\} = F_{ij}(t)$$

for $l \geqslant 1$. Thus, the transition may be considered in two stages. First, when state i is entered, the next state is chosen according to Markov transition probabilities p_{ij}. Secondly, given that the state chosen is j, the sojourn time from i to j has a distribution $F_{ij}(\cdot)$. Note that the semi-Markov model, where $F_{ij}(\cdot)$ is any distribution function, is more general than a Markov process, in which $F_{ij}(\cdot)$ is assumed to be the exponential or geometric distribution function. Suppose that there exists the density function $f_{ij}(\cdot)$. Then following the construction of the underlying probability space by Moore and Pyke [9], we have the corresponding joint density of Equation (1) as

$$\theta_{j_0} \prod_{l=1}^{m} p_{j_{l-1}j_l} f_{j_{l-1}j_l}(t_l) \tag{2}$$

with respect to the dominating product Lebesgue and counting measures.

3. ANALYSIS

Suppose that there are g values of covariates $\mathbf{Z} = (Z_1, \ldots, Z_g)'$ available for each patient and we want to incorporate this information into the probability density (2). The force of transition from state i to state j given a covariate vector $\mathbf{Z}$ is defined as

$$\lambda_{ij}(t|\mathbf{Z}) = \frac{f_{ij}(t|\mathbf{Z})}{1 - F_{ij}(t|\mathbf{Z})}.$$

We follow the proportional hazards model of Cox [2] and approximate the underlying force of transition by a step function as discussed by Kalbfleish and Prentice [5]. That is, the force of transition $\lambda_{ij}(t|\mathbf{Z})$ can be formulated as

$$\lambda_{ij}(t|\mathbf{Z}) = \lambda_{ij}^0(t) \exp(\boldsymbol{\beta}_{ij}'\mathbf{Z}), \tag{3}$$

where $\boldsymbol{\beta}_{ij}' = (\beta_{ij1}, \ldots, \beta_{ijg})$ are vectors of regression parameters measuring the effect of $\mathbf{Z}$ on the force of transition $i \to j$, and $\lambda_{ij}^0(t)$ can be expressed as

$$\lambda_{ij}^0(t) = \lambda_{ijk} \quad \text{for} \quad t_{k-1} \leqslant t < t_k, \quad 1 \leqslant k \leqslant K_{ij}, \tag{4}$$

where $t_0, t_1, \ldots, t_{K_{ij}}$ are preassigned constants. Here we choose the proper t's

and assume that within each interval the underlying forces of transition λ_{ijk} are constant. In addition, our model permits the general case where the grouping of sojourn times for different transitions may vary.

Define $\Lambda_{ij}(t|\mathbf{Z})$ as the cumulative force of transition from state i to state j. From Equations (3) and (4), we have, for $t_{k-1} \leqslant t < t_k$,

$$\Lambda_{ij}(t|\mathbf{Z}) = \exp(\boldsymbol{\beta}_{ij}'\mathbf{Z}) \cdot \left\{ \sum_{l=1}^{k-1} \lambda_{ijl}(t_l - t_{l-1}) + \lambda_{ijk}(t - t_{k-1}) \right\},$$

$$f_{ij}(t|\mathbf{Z}) = \lambda_{ijk} \exp\{\boldsymbol{\beta}_{ij}'\mathbf{Z} - \Lambda_{ij}(t|\mathbf{Z})\}. \tag{5}$$

Substituting (5) into (2), we obtain the likelihood function contributed by the individual whose transition history is as defined in (1). Since we assume that the underlying force of each transition is constant during each interval, we must decide how many intervals to construct and choose a method for grouping the sojourn times for each transition. If the sojourn times are naturally grouped with proper clinical interpretation, one need not create arbitrary intervals. Otherwise, arbitrary intervals are constructed so that the number of transitions occurring per interval is not too small.

Consider N independent patients in a clinical trial. The information on covariate vector $\mathbf{Z}_n$ $(n = 1, \ldots, N)$ and transition history for each individual are recorded. The overall likelihood function is the product of likelihood functions contributed by all individuals, and may be factored into a component for each transition:

$$L(\boldsymbol{\theta}, \mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\beta}) = \left(\prod_{i \in \mathbb{T}} \theta_i^{s_i} \right) \prod_{n=1}^{N} \prod_{i,j} \left\{ p_{ij}^{b_{ijn+}} \left(\prod_{k=1}^{K_{ij}} \lambda_{ijk}^{b_{ijnk}} \right) \right.$$

$$\left. \times \exp\left[b_{ijn+} \cdot (\boldsymbol{\beta}_{ij}'\mathbf{Z}_n) - \left(\sum_{k=1}^{K_{ij}} \lambda_{ijk} c_{ijnk} \right) \exp(\boldsymbol{\beta}_{ij}'\mathbf{Z}_n) \right] \right\}, \tag{6}$$

where the rearranged observed data are

s_i = number of patients starting in state $i \in \mathbb{T}$,

b_{ijnk} = number of transitions $i \to j$ made in kth interval by individual n during the process, $k = 1, \ldots, K_{ij}$,

$b_{ijn+} = \sum_{k=1}^{K_{ij}} b_{ijnk}$ = number of transitions $i \to j$ by individual n during the process,

c_{ijnk} = amount of time spent in the kth interval by individual n during the transition $i \to j$.

If individual n makes no transition from state i to state j, the value of each b_{ijnk} is zero and thus b_{ijn+} equals zero. If individual n makes only one

transition from state i to state j in the kth interval, then b_{ijnk} equals one and b_{ijn+} is also one, and the values of the remaining $b_{ijnk'}$ $(k' \neq k)$ are all equal to zero. When more than one, say two, transitions are made from state i to state j in the kth and k'th intervals, then $b_{ijn+} = 2$ and $b_{ijnk} = 1$, $b_{ijnk'} = 1$. A similar argument can be applied to individuals making more than two transitions from state i to state j.

The maximum likelihood estimators of multinomial-type probabilities θ_i and p_{ij} are

$$\hat{\theta} = s_i / N$$

and

$$\hat{p}_{ij} = \frac{\sum_{n=1}^{N} b_{ijn+}}{\sum_{j} \sum_{n=1}^{N} b_{ijn+}}.$$

However, there exist no closed forms for the MLEs of $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$, and we have to apply a numerical technique, e.g., the Newton-Raphson algorithm. The asymptotic properties of $(\hat{\boldsymbol{\lambda}}', \hat{\boldsymbol{\beta}}')$ can be easily derived. Since we take covariates into consideration, the sojourn times $\{T_i\}$ are independent, but they need not be identically distributed for different individuals making the same transition. Therefore, under mild regularity conditions (e.g. Hoadley [4]), the asymptotic distribution of $(\hat{\boldsymbol{\lambda}}', \hat{\boldsymbol{\beta}}')$ is a multivariate normal distribution with mean vector $(\boldsymbol{\lambda}', \boldsymbol{\beta}')$ and variance-covariance matrix equal to minus the inverse of the sample information matrix. In the Appendix, the form for the variance-covariance matrix is derived. For inference problems related to $\boldsymbol{\beta}$, that is, to make inferences about the effects of the covariate vector $\mathbf{Z}$, we may use standard techniques for testing hypothesis in multivariate normal cases.

The overall survival function is in general difficult to obtain. We will therefore consider only the construction of the survival function for the r competing risks model. In this model, we have a single initial state, say state 0, and r causes of failure, say state 1 up to state r, as the absorbing set. The overall survival function for an individual with covariate vector $\mathbf{Z}$ is estimated by

$$\hat{S}(t|\mathbf{Z}) = \sum_{j=1}^{r} \hat{p}_{0j}\left[1 - \hat{F}_{0j}(t|\mathbf{Z})\right], \tag{7}$$

where p_{0j} is the transition probability from initial state 0 to state j, and $F_{0j}(\cdot)$

is the conditional cdf of the survival time for an individual patient with the failure type j. This estimator, similar to one discussed in Peterson's paper [10], provides an alternative to those obtained by the latent time approach or by the cause-specific hazard function approach (e.g., Kalbfleisch and Prentice [6, Chapter 7]).

4. EXAMPLE

In this section, we illustrate our model and its application by analyzing the survival time of patients in a heart transplant study. The data came from the Stanford Heart Transplant Program and have been analyzed previously by Turnbull, Brown, and Hu [13], Mantel and Byar [8], Crowley and Hu [3], and Beck [1]. Turnbull et al. [13] and Mantel and Byar [8] considered the effects on the survival function of changing an individual's status from the nontransplanted to the transplanted group, treating the patient population as homogeneous (i.e., no covariates). Crowley and Hu [3] used Cox's proportional hazards model to discover the covariate values for which transplantation is likely to be beneficial. But they did not distinguish the possible different effects of covariates on transplanted patients dying from different causes. Beck [1] developed a stochastic survival model which incorporates covariate information and allows us to estimate the effects of covariates on transplanted patients dying from different causes. However, Beck used the latent time approach, which assumes independence among different stages (or causes of failure) and hence has no clear physical interpretation. With the approach of a semi-Markov model described herein, the assumption of independence is relaxed.

To evaluate the survival of heart transplantation, the state space of a semi-Markov process is constructed as follows. Two transient states are S_1 (being accepted into the program) and S_2 (receiving a new heart). Three absorbing states are R_1 (dying from rejecting the donor heart), R_2 (dying from any other causes), and R_3 (censored due to the termination of the study). Following Beck's suggestion, we include 94 patients in this analysis. The number of patients in the various transitions by the end of this study

Transition	Number	Transition	Number
$S_1 \to R_2$	28	$S_1 \to S_2 \to R_2$	12
$S_1 \to R_3$	2	$S_1 \to S_2 \to R_3$	24
$S_1 \to S_2 \to R_1$	28		

The set of covariates available in this analysis are Z_1 = age at transplant, Z_2 = previous open-heart surgery (1 = yes, 0 = no), three measures of the degree of mismatch between patient and donor (Z_3 = number of mismatches, Z_4 = measure of HLA-A2, and Z_5 = mismatch score), and Z_6 = age at accep-

tance into the program. The present analysis focuses on evaluating the covariates relating to the survival of posttransplant patients.

In the first part of this analysis, we include five covariates (Z_1, Z_2, Z_3, Z_4, and Z_5) and three possible transitions ($S_2 \to R_1$, $S_2 \to R_2$, and $S_2 \to R_3$). To decide into how many intervals to group the sojourn durations, we assign the time division points as $t_0 = 0$, $t_1 = 6$ months, $t_2 = 24$ months, $t_3 = 48$ months, and $t_4 = \infty$. We also assume that the underlying forces of transition within each time interval are constant, but they may vary between intervals. The vectors of regression coefficients corresponding to each of the covariates and the underlying forces of transition are denoted, respectively, by $\boldsymbol{\beta}'_{ij} = (\beta_{ij1}, \beta_{ij2}, \beta_{ij3}, \beta_{ij4}, \beta_{ij5})$ and $\boldsymbol{\lambda}'_{ij} = (\lambda_{ij1}, \lambda_{ij2}, \lambda_{ij3}, \lambda_{ij4})$ for the transition $i \to j$. The estimated values for p, $\boldsymbol{\lambda}$, and $\boldsymbol{\beta}$ are provided in Table 1. The estimated standard errors of $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\beta}}$ are also given in the parentheses. The results in Table 1 indicate that the measure of mismatch on HLA-A2 (Z_4) is not a significant factor in any of the three transitions. For death by rejection (i.e., $S_2 \to R_2$), age at transplant, previous surgery, and the mismatch score are important factors. These findings agree with those of Crowley and Hu [3] and Beck [1]. But in contrast to these studies, we find that the number of mismatches (Z_3) is also an important factor and that the estimated regression coefficient $\hat{\beta}_3$ is positive. This means that the larger the number of mismatches (Z_3), the less the probability of survival. This is exactly what one might anticipate. For death from other causes (i.e., $S_2 \to R_3$), previous surgery is the only significant factor. In conclusion, young age, having open-heart surgery, and a small number of mismatches are the favorable factors in prolonging survival.

For an individual with a given set of covariate values, the posttransplant survival function can be calculated from Equation (7). According to Beck, the average values of the covariates of all individuals are age at transplant, 45.6 years; number of mismatches, 2.7; and mismatch score, 1.17. For an individ-

TABLE 1

Parameter Estimators and Standard Errors in the Fitted Model

			Underlying force of transition				Regression coefficients				
Transition	n	p	λ_1	λ_2	λ_3	λ_4	β_1	β_2	β_3	β_4	β_5
$S_2 \to R_2$	12	.1875	.0820	.1672	[a]	[a]	.0388	−1.7571	[b]	[b]	.3654
			(.0132)	(.0263)			(.0038)	(.1291)			(.1041)
$S_2 \to R_1$	28	.4378	.0052	.0016	.0056	[a]	.0354	−.3491	.1644	[b]	.9400
			(.0015)	(.0004)	(.0015)		(.0047)	(.0808)	(.0276)		(.1212)
$S_2 \to R_3$	24	.3750	.0356	.0318	.0656	.0730	[b]	−.3514	[b]	[b]	[b]
			(.0040)	(.0033)	(.0067)	(.0104)		(.0589)			

[a] The MLE is undefined since no transitions occur in this interval.

[b] The MLE is not significant at $\alpha = 0.05$ and thus not included.

TABLE 2

Survival Function for an Individual with Average Covariate Values and with Previous Surgery

Time (months)	Survival probability
1	.9304
3	.8667
5	.7063
10	.5496
15	.4600
20	.3975
25	.2817
30	.1974
35	.1396

ual with previous surgery (i.e., $Z_1 = 1$) and the average covariate values, the estimated posttransplant survival function for is shown in Table 2. The mid-life time in this case is about 12.5 months.

In conclusion, with a semi-Markov model, we could study the parameters associated with the marginal subsurvival distributions for the mutually exclusive states and the overall survival function as well. An alternate approach may be obtained by application of separate and independent Cox models for each pair of adjacent states. However, the latter approach is a latent time approach with a strong assumption of independence.

APPENDIX. DERIVATION OF ASYMPTOTIC VARIANCE-COVARIANCE MATRIX $\Sigma_{(\hat{\lambda},\hat{\beta})}$

The reduced log likelihood function after eliminating terms of θ_i and p_{ij} from Equation (6) is

$$\phi = \sum_i \sum_j \sum_n \left\{ \sum_k (b_{ijnk} \ln \lambda_{ijk}) + \left(\sum_k b_{ijnk} \right) (\boldsymbol{\beta}_{ij}' \mathbf{Z}_n) - \left(\sum_k \lambda_{ijk} c_{ijnk} \right) \exp(\boldsymbol{\beta}_{ij}' \mathbf{Z}_n) \right\}.$$

The first and second derivatives of ϕ with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ are as follows:

$$\dot{\phi}_{\lambda_{ijk}} = \frac{\sum_n b_{ijnk}}{\lambda_{ijk}} - \sum_n \{ c_{ijnk} \exp(\boldsymbol{\beta}_{ij}' \mathbf{Z}_n) \},$$

$$\dot{\phi}_{\beta_{ijl}} = \left[\sum_n \left(\sum_k b_{ijnk} \right) Z_{nl} \right] - \sum_n \left\{ \left(\sum_k \lambda_{ijk} c_{ijnk} \right) Z_{nl} \exp(\boldsymbol{\beta}_{ij}' \mathbf{Z}_n) \right\}, \tag{A.1}$$

and

$$\ddot{\phi}_{\lambda_{ijk}\lambda_{ijk'}} = \begin{cases} -(\Sigma_n b_{ijnk})/\lambda_{ijk}^2 & \text{for } k' = k, \\ 0 & \text{otherwise,} \end{cases}$$

$$\ddot{\phi}_{\lambda_{ijk}\beta_{ijl}} = -\sum_n \{c_{ijnk} Z_{nl} \exp(\boldsymbol{\beta}_{ij}' \mathbf{Z}_n)\}, \tag{A.2}$$

$$\ddot{\phi}_{\beta_{ijl}\beta_{ijl'}} = -\sum_n \left\{ \left(\sum_k \lambda_{ijk} c_{ijnk} \right) Z_{nl} Z_{nl'} \exp(\boldsymbol{\beta}_{ij}' \mathbf{Z}_n) \right\}.$$

The MLEs $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\beta}}$ may be obtained by solving the equations (A.1) set equal to zero.

The asymptotic variance-covariance matrix of $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}})$ is equal to minus the inverse of the sample information matrix. Denote

$$A_N = \left(-\ddot{\phi}_{\lambda_{ijk}\lambda_{ijk'}} \right),$$

$$B_N = \left(-\ddot{\phi}_{\beta_{ijl}\beta_{ijl'}} \right),$$

$$C_N = \left(-\ddot{\phi}_{\lambda_{ijk}\beta_{ijl}} \right),$$

where $\ddot{\phi}$'s are given in (A.2). Then by inverting the matrix, we have

$$\Sigma_{(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\beta}})} \simeq \begin{bmatrix} D_N^{-1} & -A_N^{-1} C_N E_N^{-1} \\ -B_N^{-1} C_N' D_N^{-1} & E_N^{-1} \end{bmatrix},$$

where

$$D_N = A_N - C_N B_N^{-1} C_N',$$

$$E_N = B_N - C_N' A_N^{-1} C_N.$$

The author wishes to acknowledge helpful comments made by the referees and the editor.

REFERENCES

1 G. J. Beck, Stochastic survival models with competing risk and covariates, *Biometrics* 34:427–438 (1979).

2 D. R. Cox, Regression models and life tables (with discussion), *Roy. Statist. Soc. Ser. B* 34:187–220 (1972).

3 J. Crowley and M. Hu, The covariance analysis of heart transplant data, *J. Amer. Statist. Assoc.* 72:27–36 (1977).

4 B. Hoadley, Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case, *Ann. Math. Statist.* 42:1977–1991 (1971).
5 J. D. Kalbfleish and R. L. Prentice, Marginal likelihood based on Cox's regression and life model, *Biometrika* 60:267–278 (1973).
6 J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
7 S. W. Lagakos, C. J. Sommer, and M. Zelen, Semi-Markov models for partially censored data, *Biometrika* 65:311–318 (1978).
8 N. Mantel and D. P. Bayar, Evaluation of response-time data involving transient states: An illustration using heart-transplant data, *J. Amer. Statist. Assoc.* 69:81–86 (1974).
9 E. H. Moore and R. Pyke, Estimation of the transient distributions of a Markov renewal process, *Ann. Inst. Statist. Math.* 20:411–424 (1968).
10 A. V. Peterson, Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions, *J. Amer. Statist. Assoc.* 72:854–858 (1977).
11 R. L. Prentice, J. D. Kalbfleisch, A. V. Peterson, N. Flournoy, V. T. Farewell, and N. Breslow, The analysis of failure time data in the presence of competing risks, *Biometrics* 34:541–554 (1978).
12 R. Pyke, Markov renewal precesses: Definitions and preliminary properties, *Ann. Math. Statist.* 32:1231–1242 (1961).
13 B. W. Turnbull, B. W. Brown, and M. Hu, Survivorship analysis of heart transplant data, *J. Amer. Statist. Assoc.* 69:74–80 (1974).
14 H. H. Weiss and M. Zelen, A semi-Markov model for clinical trials, *J. Appl. Probab.* 2:269–285 (1965).