AN APPLICATION OF EM BASED MULTISTATE MODEL IN PROSTATE CARCINOMA DATA

S. Bae[1,2], K.P. Singh[1], A.A. Bartolucci[1], H. Ragde[3,4]

1. Department of Biostatistics, School of Public Health
University of Alabama at Birmingham
Birmingham, Alabama 35294-0022, U.S.A.
2. Division of Preventive Medicine, School of Medicine
University of Alabama at Birmingham
Birmingham, Alabama 35294-0022, U.S.A.
3. Pacific Northwest Cancer Foundation/Northwest Hospital,
120 Northgate Plaza, Suite 200
Seattle, WA 98125, U.S.A.
4. Department of Urology, University of Washington,
Seattle, WA 98195, U.S.A.

For the follow up studies we observe the repeated observation of outcome and prognostic factors over time. The study of a subject over time may show changes from one outcome state to another and the histories of a group of individuals may include the partially censored data or missing data. In such a case, an Expectation-Maximization (EM) based parameter estimation is an intuitive approach. The transition from one state to another can be categorized into three distinct types: progression, regression, and absorbing transition. In this paper a method is exploited for estimating parameters of the model for reverse and repeated transition developed by Kay (1982) and further extended by Islam and Singh (1992) and Singh *et. al.* (1999). For estimating parameters in the multistate model, the EM method was utilized. We apply the model to follow up data on Prostate Carcinoma from the Pacific Northwest Cancer Foundation/Northwest Hospital.

1. Introduction

Survival analysis is a series of statistical approaches for data analysis for which the variable of interest is time until an event occurs. By time, we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs. By event, we mean death, disease onset, patient relapse, or any designated experience of interest that may happen to an individual. In general, we can think of the time as the survival time and the event as the failure time. Survival analysis is distinguished from most other analysis by the incorporation of censoring; that is, some individuals' failure times are unknown due to withdrawal or survival beyond the end time of the study. Therefore, pairs of observed time and indicator functions of failure are recorded in survival data.

In typical survival models the intermediate transitions are not accounted for and only the aggregate time to death is analyzed (Figure 1A). For example, in a survival model of two transition states and one absorbing state, for State 2 patients the transition to State 3 characterized by α_{23} is usually ignored and so are the prognostic factors affecting the intermediate $2 \rightarrow 3$ transition. The death rate used in survival models is a mixture of α_{13} and α_{23}, the mixture probabilities being the conditional probabilities of being in State 1 or State 2 given that the patient started in State 1 at time 0 and survived up to time t (Andersen 1988). We define a two-state model as one where only two states are considered: the initial state and the outcome state. The outcome state is either death or a progressive stage, and no intermediate transitions are considered. In contrast, competing risks data arise from various censoring mechanisms. A competing risk model may have more than two absorbing states (Figure 1B), three absorbing states. For example, when analyzing the risk factors for the onset of stroke many of the subjects may die from a heart attack prior to having a stroke. Those individuals who die from a heart attack are considered censored from the analysis due to this competing risk. On the other hand, a multistate model is defined as one that incorporates all transitions into a comprehensive model (Figure 1C).

The nonparametric likelihood method for a multistate stochastic process has been developed by Lagakos, Somer, Zelen (1978). They assumed the underlying semi-Markov model as an embedded Markov chain,

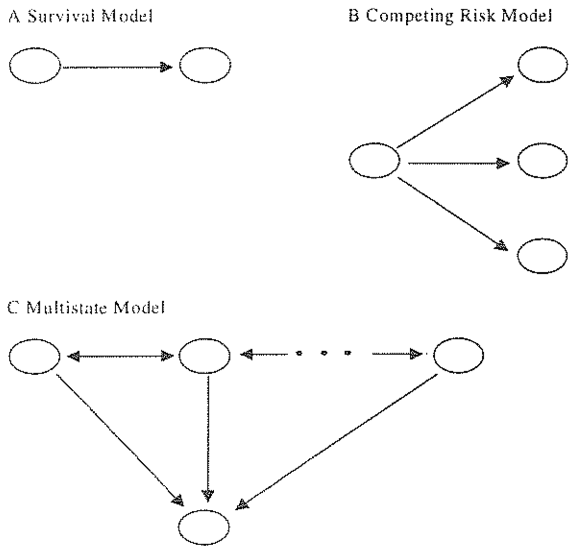

Figure 1. Various disease state modelling.

and transition times are independent and depend only on adjoining states. A recharacterization of their model is shown by Dinse and Larson (1986) that can be used to express the nonparametric likelihood estimators as a function of cause-specific hazard estimators and the product-limit estimator. They have not incorporated any covariates in the model. Voelkel and Crowley (1984) used a semi-Markov specification for forward going proportional hazards. Aalen et. al. (1980) used Markov chain models for analyzing interaction between life history events. Kay (1982) showed an extension of the proportional hazards model for several transient states. His model also considered a hierarchial approach. The partial likelihood factor for any transition in Kay's model is identical to the partial likelihood for Cox's (1972) model except for the definition of the risk sets. Beck (1979) developed a stochastic survival model which incorporates two health states and several absorbing states. Beck's model does not consider reverse transition among transient states directly.

This paper discusses a general k state multistate model in which the exact transition times are not observed. Of particular, relevance is the extension of the relation between the multistate models and survival analysis functions. Multistate models represent a generalization of parametric models in survival analysis to the analysis of data concerning multiple events.

Multistate models are extended to include covariates in the transition intensities as in proportional hazard models. An important application of this model is discussed and analyzed. Data from a longitudinal study in prostate cancer from the Pacific Northwest Cancer Foundation are used to find relations of markers for prostate cancers and to describe the natural course of prostate carcinoma.

2. Mulitistate Models

The multistate models have become important tools to describe and help understand the progression and regression processes of important multistate diseases, such as cancer, HIV infection, AIDS, diabetes, and many other chronic diseases. These models have been used and discussed by many authors, including Andersen (1988), Gail (1981), Kay (1982), and Prentice and Williams (1981). Especially, in the area of molecular biology, research has been done on possible markers for the transition from stable states to the accelerated phase, the irreversible (absorbing) state of a disease, or both, thus, describing the natural course of diseases.

The multistate model is an extension of the basic Cox regression model for right censored survival data (Cox 1972). The multistate model is the study of the occurrence rate of the several types of events that individuals may experience in their lifetimes (Andersen & Borgan 1985; Clayton 1988; Keiding et al. 1989). The simplest form of this model is the competing risks model (Andersen & Borgan, 1985; Cox & Oakes, 1984; Prentice et al., 1978) where interest focuses not only on death but also on the causes of death.

There are different types of events that could be labeled, $i = 1, \ldots I$, and for which models are most conveniently specified by means of the intensities $\alpha_{ij}(t)$ of occurrence of the events of the different types. Following the instantaneous failure rate convention, this intensity has the interpretation that $\alpha_{ij}(t)\Delta t$ when $\Delta t > 0$ is small is approximately the conditional probability that individual j experiences an event of type i in the interval from t to $t + \Delta t$ given that entire past up until just before time t if individual I is at risk for an event of type I at that time. A Cox type regression model can now be considered by specifying the intensity as

$$\alpha_{ij}(t) = \alpha_{i0}(t)\exp(\beta_i^T Z_j(t))$$

or

$$\alpha_{ij}(t) = \alpha_{i0}(t)\exp(\beta^T Z_{ij}(t))$$

where the covariates of interest may differ between the different types of events. In fact, the model can also be written by appropriately defining type specific covariates (Andersen & Borgan, 1985). Large sample properties of multistate models have been derived by authors, including Andersen and Gill (1982) and Andersen and Borgan (1985). Some examples of the use of multistate models in medical research have been given by Houggard and Madsen (1985), and Andersen (1988). Klein et. al. (1984) used a three-state semi-Markov model in a study of patients with chronic myelogenous leukemia, to analyze the effect of elevated blood levels of adenosine deaminase as a marker for transition from stable disease to blast crisis and then to death.

3. Product Limit (PL) Based Multistate Model

The PL method for one transient state and one absorbing state (Kaplan & Meier 1958; Meier 1977) can be extended for multiple causes of decrement for transitions from one transient state to several absorbing states on the basis of simpler assumptions. In the PL method censoring one transient state and one absorbing state are considered. However, in many practical situations, one has to deal with more than one absorbing state. For instance, in morbidity studies, there are a number of causes of decrement. A simple generalization of the PL method can be shown for a single transient state and r absorbing states where absorbing states represent the causes of decrement from a single healthy state or transient state. The k distinct failure times are $t_1 < t_2 < ... < t_k$ in a sample of size n_j. Let us define α_j^u as the hazard component at time t_j for cause of decrement u (u = 1, 2 ,..., r; j = 1, 2, ... , k), d_j^u as the number of failures due to cause u at time t_j, c_j as the number of observations censored during the time interval $[t_j, t_{j+1})$, n_j as the number of individuals at risk at a time just prior to t_j. To obtain the likelihood function we have assumed that (a) the failures occur at discrete points in time, t_j, j = 1, 2, ..., k, to n_j individuals exposed to the risk of failure due to causes of failure u (u = 1, 2, ..., r); (b) no individual could have more than one failure at t_j; and (c) censoring occurs at the end of the interval $[t_j, t_{j+1})$, such that the number of individuals censored during the interval can be subtracted from the population exposed to failure at time t_j to obtain the number of individuals exposed to the risk of failure at the subsequent point in time, t_{j+1}. In other words, at the end of the interval $[t_j, t_{j+1})$, we

obtain a new set of exposed population members who survived failures or censoring during the interval. Hence, the probabilities for competing risk at time t_j can be expressed as:

$$\alpha_j^u = Pr\,(T = t_j,\ U = u \mid T \geq t_j),\ t_j \in T,\ u \in U;$$

and the probability of no event at time t_j is

$$(1 - \sum_{u=1}^{r} \alpha_j^u).$$

Thus the exposed population at time t_{j+1} is reduced to

$$n_{j+1} = n_j - \sum_{u=1}^{r} d_j^u - c_j.$$

Under these assumptions, the likelihood can be obtained. Using the multinomial distribution (Islam & Singh, 1992 ; Singh et al., 1999) we obtain the maximum likelihood estimators

$$\hat{\alpha}_j^u = \frac{d_j^u}{n_j}, \qquad u = 1,2,\ldots,r.$$

Using the relationship between $F(t)$ and competing risk failure rates λ_j^u (Kalbfleisch & Prentice 1980), the estimate for the survival function, $F(t)$, is obtained as follows:

$$\hat{F}(t) = \prod_{t_j<t}\prod_{u=1}^{r}(1 - \hat{\alpha}_j^u) = \prod_{t_j<t}\prod_{u=1}^{r}(\frac{n_j - d_j^u}{n_j})$$

Kay (1986) proposed a methodology to fit a general k disease state Markov model in continuous time with application to the analysis of cancer markers in survival studies. Longini et al. (1989) used the same model to describe the distribution of the incubation period for AIDS patients. Kalbfeisch and Lawless (1985) introduced a continuous-time Markov model to analyze panel data, and Kalbfleisch et. al. (1983) proposed methods to estimate the parameters of this model from aggregate data. Beck (1979) developed a stochastic survival model that incorporates two transient states and several absorbing states. However, Beck did not extend this model for reverse transi-tions. Islam and Singh (1992) extended Beck's model for transitions as well as reverse transitions (Singh et al. 1999).

4. APPLICATION

We analyzed the data set collected from the Pacific Northwest Cancer Foundation and Northwest Hospital where 144 prostate carcinoma patients who went through Iodine-125 radio nuclides (Brachytherapy) or Brachytherapy combined with radiation therapy. These patients were followed up yearly for 10 years after the treatment. The goal of the analysis was to determine whether the use of radiation therapy helped lower the risk or increased the time to the absorbing state as well as from which transient the patients have lower risk to the absorbing state. Prostate specific antigen molecule (PSA) is believed to leak from the prostatic ductal system into the prostatic stroma and then into the blood stream via capillaries and lymphatics. Unlike traditional tumor markers, PSA is not found in larger amounts in tumor cells as compared with healthy tissue. In fact, the opposite is true: malignant prostate tissue actually produces less PSA than normal prostate epithelial cells and benign prostatic adenomatous tissue (Papsidero et al. 1981; Qui et al. 1990). When monitored serially after treatment, serum PSA is considered the most universally verified and validated method to determine disease-free survival whether the treatment is by radiation or by surgery (Kaplan et. al. 1993; Partin et al. 1993; Ravery et al. 1994; Ritter et al. 1992; Zagars 1992; Zagars & Pollack 1995).

Of the 144 patients, 50 had the events of interest. Events of interest were different failure types: bone scan failure, biopsy failure, lost to follow up, and PSA failure. PSA measurements in patients treated with external beam irradiation were followed from 8 to 18 years. It has been postulated that the optimum relapse-free value after radiation therapy should be 0 to 1.0 ng/mL (Pisansky et al. 1993). A patient could start in either State 0 (PSA <= 1.0 ng/mL) or State 1 (PSA > 1.0 ng/mL). State 0 and State 1 are the transient states. Patients who had events of interest are considered to be in State 2, that is, the absorbing state. The distribution of the patients' transition states are given in Table 1. Note that more than 50 % of the patients had reduced PSA level: 1 -> 0 transition.

Table 1.
Follow-Up Transition History for the Prostate Cancer Study

Transition history	N
0	4
0 -> 1 -> 0	2
1 -> 0	79
1 -> 0 -> 1 -> 0	8
1 -> 0 -> 1 -> 2	10
1 -> 0 -> 2	17
1 -> 2	22
1 -> 0 -> 1 -> 0 -> 1 -> 0 -> 1 -> 0 ->2	1
0 -> 1 -> 0 -> 1 -> 0 -> 1 -> 0	1
Total	144

Note. The last two transition histories have been deleted from this analysis due to unusual transitions from the rest of the patients. N denotes the number of patients for a given history.

Parameter Estimation for EM Multistate Model: To find the initial estimates of the method

$$\alpha_{ij}^{(0)} = \frac{n_{ij}}{T_i}$$

given in Kay (1986) is used: $\alpha_{i2}^{(0)} = \dfrac{n_{i2}}{T_i}$

Using the above estimates, the EM-based multistate prostate carcinoma algorithm is given in Table 2. The algorithm consists of eight steps to estimate the α_{ij} parameters for the intensity matrix and the calculation of the transition probability matrix using the intensity matrix.

Table 2.
EM Multistate Algorithm for the Prostate Carcinoma Data

1. Find the initial value of the parameters ($\hat{\alpha}_{ij}$)

2. Find $E(T_{ij} | \alpha_{ij}, x_{ij})$: E-step
3. Find MLE of $\hat{\alpha}_{ij}$ and update the value of the $\hat{\alpha}_{ij}$: M-Step
4. Calculate $L_k(\alpha_{ij}; T_{ij}, x_{ij}) \ where \ k = 0, 1, ...$
5. Compare L_k, L_{k+1} iterate Steps 2 to 4 until it converges (0.00001 criterion)
6. Form transition matrix Q based on Steps 1 to 5
7. Find Eigenvalues and Eigenvectors of the Q matrix
8. Find probability matrix *P* based on Steps 6-7

Note. EM denotes Expectation and Maximization, MLE denotes Maximum Likelihood Estimator.

The EM part of the Table 2 converged after 50 iterations, and the estimates are shown in Table 3.

Note that the median survival time is around 10.64 years using the KM analysis. We find more valuable insight of the data using the multistate analysis. The transition intensities can be interpreted as the number of transitions in a constant period of time. The transition rate from State 1 to State 0 is about 40 times more likely than from State 0 to State 1. Relatively speaking, having entered State 0 there is a smaller chance of returning to State 1. Also, the transition matrix illustrates that a patient who leaves State 0 has 1.1 times chance to progress to State 1. Similarly, a patient who leaves State 1 has 3.9 times greater chance to regress to State 0. Furthermore, the relative risk of proceeding from State 1 to absorbing compared to that of State 0 is 11.33, which indicates that increases in PSA level are strongly associated with increase in the risk of failure.

Table 3.
EM Estimates and Log Likelihood for the Prostate Carcinoma Data

	Parameters			
Iter	$\hat{\alpha}_{01}$	$\hat{\alpha}_{02}$	$\hat{\alpha}_{10}$	$\hat{\alpha}_{12}$
Init	0.0308	0.0275	0.4269	0.1076
10	0.0621	0.0556	2.2398	0.5645
20	0.0613	0.0549	2.4431	0.6157
30	0.0612	0.0548	2.4635	0.6208
40	0.0612	0.0548	2.4656	0.6214
50	0.0612	0.0548	2.4658	0.6214
60	0.0612	0.0548	2.4658	0.6214
90	0.0612	0.0548	2.4658	0.6214

Iter means Iteration and Init means Initial.

The value of the log-likelihood is -155.709.

5. References

Andersen, P. K. (1988). Multistate models in survival analysis: A study of nephropathy and mortality. *Statistics in Medicine*, 10, 1931- 1941.

Andersen, P. K., & Borgan, O. (1985). Counting process models for life history data: A review (with discussion). *Scandinavian Journal of Statistics*, 12, 97-158.

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, 10, 1100-1120.

Beck G. J. (1979). Stochastic survival models with competing risks and covariates. *Biometrics*, 35, 427-438.

Clayton, D. (1988). The analysis of event history data: A review of progress and outstanding problems. *Statistics in Medicine*, 7, 819-841.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187-220.

Demster, A. P., Laird, N. M., & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society , Series B*, 39, 1-38.

Dinse, G. E., & Larson, M. G. (1986). A note on semi-Markov models for partially censored data. *Biometrika*, 73, 379-386.

Gail M. H. (1981) evaluating serial cancer marker studies in patients at risk of Recurrent disease. *Biometrics*, 37(1), 67-78.

Houggard, P., & Madsen, E. B. (1985) dynamic evaluation of short-term prognosis of myocardial. *Statistics in Medicine*, 4, 29-38.

Islam, M. A. and Singh, K. P. (1992) multistate survival models for partially censored data. *Environmetrics*, 3(2), 223-234.

Kalbfleisch, J. D., & Lawless, J. F. (1985). The analysis of panel data under a markov assumption. *Journal of the American Statistical Association*, 80, 832-871.

Kalbfleisch, J. D. & Prentice, R. L. (1980). *The statistical analysis of failure time data.* New York: John Wiley and Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimates from incomplete observations. *Journal of the American Statistical Association*, 53, 453-457.

Kaplan, I. D., Cox, R. S., & Bagshaw, M. A. (1993). Prostate-specific antigen after external beam radiotherapy for prostate cancer: Follow-up. *Journal of Urology*, 149, 519-522.

Kay, R. (1982). The analysis of transition times in multistate stochastic process using proportional hazard regression models. *Communications in Statistics A*, 11, 1743-1756.

Keiding, N., & Andersen, P. K. (1989). Nonparametric estimation of transition intensities and transition probabilities: A case study of two-state Markov process. *Applied Statistics*, 38, 319-329.

Klein, J. P., Klotz, J. H., & Grever M. R. (1984). A biological marker model for predicting disease transitions. *Biometrics*, 40, 929-936.

Longini, I. M., Clark, W. S., Byers, R. H., Ward, J. W., Darrow, W. W., Lemp, G. F., & Hethcote, H. W. (1989). Statistical analysis of the stages of HIV infection using a markov model. *Statistics in Medicine*, 8, 831-843.

Meier, P. (1977). Estimation of a distribution function from incomplete observations. In J. Gani(Ed). *In Perspectives in Probability and Statistics* (pp. 67-87). New York: Academic Press.

Papsidero, L. D., Kuriyama, M., Wang, M. C., Horoszewicz, J., Leong, S. S., Valenzuela, L., Murphy, G. P., & Chu, T. M. (1981). Prostate antigen: A marker for human prostate epithelial cells. *Journal of National Cancer Institute*, 66, 37-42.

Partin, A. W., Pound, C. R., Clemens, J. Q., Epstein, J. I., & Walsh, P. C. (1993). Serum PSA after anatomic radical prostatectomy. The Johns Hopkins experience after 10 years. *Urologic Clinics of North America*, 20, 713-725.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., & Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34, 541-554.

Prentice, R. L., & Williams, B. J. (1981). On the regression analysis of multivariate failure time data. *Biometrics*, 68, 373-379.

Singh, K. P., Chowdhury, R. I., Bae, S., Islam, M. A, Bartolucci, A. A., & Warsono, W. (In Press). Estimation of multistate proportional hazards model. *Environmental International.*

Voelkel, J. G., & Crowley, J. (1984). Nonparametric inference for a class of semi- Markov processes with censored observations. *The Annals of Statistics*, 12, 142-160.