Biometrika (1992), **79**, 3, *pp.* 513–22
Printed in Great Britain

Nonparametric tests of the Markov model for survival data

BY MICHAEL P. JONES
Department of Preventive Medicine, University of Iowa, Iowa City, Iowa 52242, U.S.A.
AND JOHN CROWLEY
Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, Washington 98104, U.S.A.

SUMMARY

Jones & Crowley (1989) introduced a general class of nonparametric statistics for right-censored survival data that includes many of the existing procedures involving a single covariate. This paper considers this class in the framework of a Markov model and suggests some tests especially appropriate for the Markov setting. Application to left-truncated and right-censored data is also given. Small sample properties and asymptotic relative efficiencies for several statistics of this class are investigated for both relative and excess risk models for the hazard.

Some key words: Censored data; Excess risk; Generalized log rank test; Kendall rank correlation; Markov model; Relative risk; Truncated data.

1. INTRODUCTION

Data from survival studies typically consist of time to failure or some other event, an indicator for censoring, and covariate information. Time to failure depends not only on the usually well-defined failure endpoint, but also on the beginning point from which time is measured. For example, in a study of the risk of breast cancer after benign breast disease, time to breast cancer can be measured from birth or from the onset of benign breast disease. We consider here the problem in which there is a waiting time W until some event, say event 1, at which time the individual is considered to become at risk for the failure endpoint. This person is then followed until failure, loss due to a competing risk, or termination of study. This setting is conveniently described by a state space diagram. At time $t=0$, N individuals enter the waiting state S_0. Upon the occurrence of event 1 a subject makes a transition to an intermediate state S_1, which for future reference occurs at a rate $\theta(t)$. Subjects in state S_1 who fail enter the failure state D at a rate $\lambda_{S_1}(t, W)$. Of course, some individuals may not experience event 1 and may either remain in S_0, move to the competing state C at a rate $\psi_{S_0}(t)$, or proceed to D at rate $\lambda_{S_0}(t)$. Likewise, some individuals in S_1 may remain in S_1 or move to state C at rate $\psi_{S_1}(t, W)$. The hypothesis of interest is whether the transition rate from state S_1 to the failure state D depends on the time W in the waiting state S_0.

Two models relevant to this problem are the Markov and semi-Markov models. A common approach in clinical settings is the semi-Markov model, in which each individual's time axis is realigned so as to start at entry to the risk state S_1. The Markov model, common in epidemiology, lacks this renewal flavour by measuring time since entering state S_0. The semi-Markov model states that $\lambda_{S_1}(t, w) = \lambda_{S_1}(t-w)$ and the Markov

model states that $\lambda_{S_1}(t, w) = \lambda_{S_1}(t)$. The hypothesis of interest in this paper is that of the Markov model, that is, of whether $\lambda_{S_1}(t, w)$ is only a function of t.

One application of the Markov model is to left-truncated, right-censored data. Suppose that individuals are observed only after they enter state S_1. Prior to that time their existence is unknown. For an S_1 individual suppose the waiting time W can be determined retrospectively. If the time $\tilde{T}$ from state S_0 to failure state D cannot be observed whenever $\tilde{T} < W$, then $\tilde{T}$ is said to be left-truncated. An hypothesis of interest is whether $\tilde{T}$ is independent of W in the region $\tilde{T} \geqslant W$. This hypothesis is equivalent to the Markov hypothesis stated above in terms of the hazard λ_{S_1}. Tsai (1990) gives further discussion of this hypothesis for left-truncated data.

In this paper we consider tests of the Markov model for the setting above. In particular, § 2 considers a general class of nonparametric tests of the Markov model, while § 3 contains some examples. Section 4 investigates the Pitman efficiencies of some members of this class and introduces new tests based on maximizing the efficiency. In § 5 small sample properties of the tests are studied by Monte Carlo methods. All this is done for both relative and excess risk models for the transition rate from S_1 to D.

2. THE GENERAL CLASS OF STATISTICS

Jones & Crowley (1989) introduced a general class of nonparametric survival analysis tests involving a single possibly time-dependent covariate. In this section this class will be applied to the Markov model in which the waiting time W is the single covariate. We assume the general random censorship model throughout. Censoring occurs both because of transitions to the competing state C and because of incomplete histories.

Let us first fix some notation. Let $t_1 < \ldots < t_k$ be the ordered distinct failure times. Define an at-risk indicator $Y_j(t)$ to be 1 if the jth individual is in state S_1 at time t and 0 otherwise, and an observed failure indicator $J_j(t)$ to be 1 if the jth individual is in S_1 and is observed to fail at time t and 0 otherwise. Let $n(t) = \Sigma\, Y_j(t)$, the number at risk at time t, and $d(t) = \Sigma\, J_j(t)$, the number in S_1 to fail at time t. Note that $n(t)$ fluctuates up and down depending on the immigration rate from S_0 to S_1 and the failure and censoring rates from S_1 to states D and C, respectively. Let W_j be the waiting time for the jth person; individuals not entering S_1 will not have a value defined for W. Define the label $Z_j(t)$ to be some function of W_j, using only information from the study prior to time t.

One approach to testing the Markov model is to assume that the risk of failing at time t for an individual in state S_1 with waiting time w is given by $\lambda_{S_1}(t, w) = \lambda_{S_1}(t) \exp(w\beta)$, and testing $\beta = 0$ using the Cox partial likelihood score test. This statistic can be written as

$$\sum_{i=1}^{k} \sum_{j=1}^{N} J_j(t_i) \left\{ W_j - n^{-1}(t_i) \sum_{j=1}^{N} Y_j(t_i) W_j \right\},$$

a form which allows for ties. The general class of statistics proposed by Jones & Crowley (1989) is given by

$$T(q, Z) = \sum_{i=1}^{k} q(t_i) \sum_{j=1}^{N} J_j(t_i) \{ Z_j(t_i) - \bar{Z}(t_i) \},$$

where $q(t)$ is a weight function at time t, N is the total number of subjects who entered state S_0 and $\bar{Z}(t) = \Sigma\, Y_j(t) Z_j(t) / n(t)$ is the average label of those at risk at time t. The

Cox partial likelihood score test is thus $T(1, W)$. The proposed variance estimator for $T(q, Z)$ is

$$V(q, Z) = \sum_{i=1}^{k} q^2(t_i) d^*(t_i) \sigma_{zz}(t_i),$$

where $d^*(t) = d(t)\{n(t) - d(t)\}/\{n(t) - 1\}$, and

$$\sigma_{zz}(t) = n^{-1}(t) \sum_{j=1}^{N} Y_j(t)\{Z_j(t) - \bar{Z}(t)\}^2$$

is the sample variance of the labels of those at risk at t. When there are no tied failure times $d^*(t) \equiv 1$.

The only difference in $T(q, Z)$ and $V(q, Z)$ between the semi-Markov and Markov models is the composition of the risk set defined by the Y_j. Jones & Crowley (1990) showed using martingale theory that under the null hypothesis and regularity conditions, V is consistent for var (T), and that $T/\sqrt{V}$ is asymptotically a standard normal random variable for a general class of weight functions q and label functions Z. Although the theorem of Jones & Crowley (1990) is stated for the semi-Markov case, it applies to the Markov case without change. The failure time distribution need not be continuous, just as the distribution of waiting times need not be continuous. One regularity condition worth noting here is that the limiting proportion at risk in S_1 must be bounded away from zero over all subintervals of the study period.

Several special cases of T will be considered here, constructed by varying the weight function $q(t)$ and the label $Z_j(t)$. The most obvious choice of label is $Z_j(t) = W_j$. In this case $T(1, W)$ is equivalent to the Cox (1972) score test, as noted. A more robust choice for Z_j might be $R_j(t)/n(t)$, where $R_j(t)$ is the rank of W_j among those at risk $\{l\colon Y_l(t) = 1\}$. This motivates the generalized log rank statistic of Jones & Crowley (1989), defined by $T(1, R(t)/n(t))$, and later derived by Jones (1991) as a score test from a pseudo-partial likelihood. Use of the weight function $q(t) = n(t)/N$ along with the label $R_j(t)/n(t)$ gives the statistic $T(n(t)/N, R(t)/n(t))$, which is equivalent to the Tsai (1990) modification to the Kendall rank statistic for censored data. The Kendall rank correlation is usually defined in terms of U statistics. Its U-statistic formulation is given as equation (A·1) in Appendix 1. Many other label functions are possible, including O'Brien's (1978) normal scores and logit ranks.

Inspection of the form of the Cox score statistic $T(1, W)$ reveals that later events are likely to receive greater weight because of longer waiting times W; by analogy time-weighted versions of the generalized log rank and Kendall statistics might have promise. Such variations are investigated in §§ 4 and 5.

3. Examples

Two examples are considered here. First, let us examine the Channing House retirement community survival data for men given by Hyde (1977). These data were also used by Tsai (1990) as an example of left truncation. The waiting time or truncation time is the age at entry into the community. Age when last seen and an indicator of death are the other two variables. Of the 97 men who entered the community, 46 died. Tsai reported a Kendall statistic of 2·02. Tsai's U-statistic version of the Kendall statistic given in (A·1) and our version given as $T(n(t)/N, R(t)/n(t))$ handle tied failure times a little differently. Probably due to the existence of a few tied failure times we compute the Kendall statistic

to be 1·97. The other standardized test statistics are 1·29 for the Cox score test $T(1, W)$ and 1·38 for the generalized log rank test $T(1, R(t)/n(t))$. Only the Kendall statistic gives evidence that the 'independence' between entry time and survival time is questionable. This is due to the $n(t)/N$ weighting scheme used by the Kendall statistic. The number at risk $n(t)$ in these data rises quickly at first, is fairly constant for quite a while, then very gradually decreases.

The second example is contrived to emphasize a potentially undesirable feature of the Kendall statistic. Prentice & Marek (1979) pointed out that statistics which use the number of individuals at risk as a weight function are very sensitive to the censoring distribution. In the Markov setting the number at risk is a function not only of the censoring distribution but also the immigration rate $\theta(t)$ from S_0 to S_1. Table 1 contains a data set in which immigration is very heavy during the final stage of a study. The Cox score statistic is 2·676, the generalized log rank is 2·545 and the Kendall is 0·567. The Kendall statistic was unable to detect the obvious trend because too much weight was placed on the last two failure times.

Table 1. *Data set for which the Kendall statistic does poorly*

Failure time	Number at risk	W of failing individual	Sample mean of W	Sample variance of W	Rank of W for failing individual
2	4	1	0·50	0·25	4
3	8	2	0·75	0·35	7
5	12	4	1·25	0·60	11
6	16	3	1·50	1	13
9	20	8	3·00	8	15
13	24	8	4·50	7·5	20
15	28	14	7·50	9	24
16	100	8	8·00	16	50
17	150	8	8·50	16	72

4. ASYMPTOTIC RELATIVE EFFICIENCIES

In this section we shall investigate the Pitman asymptotic relative efficiencies of several members of the $T(q, Z)$ family under a sequence of contiguous hazard alternatives

$$H_{ca}\colon \lambda_{S_1}^N(t, w) = \lambda_1(t)\{\alpha_1(t) + \beta N^{-\frac{1}{2}} w + O(N^{-1})\}.$$

When $\alpha_1(t) = 1$, this is a relative risk model which incorporates the special case $\lambda_1(t) \exp(\beta N^{-\frac{1}{2}} w)$; when $\lambda_1(t) = 1$, it is an excess risk model. Under H_{ca} and regularity conditions, Jones & Crowley (1990, § 4) showed that the statistic $T(q, Z)/V^{\frac{1}{2}}(q, Z)$, using only information gathered during the interval $[0, t]$ within the study period, converges in distribution to a $N(\mu(t)/\sigma(t), 1)$ random variable, where $\mu(t)$ and $\sigma^2(t)$ are defined as follows for two special cases of the label:

Label	$\mu(t)$	$\sigma^2(t)$
W	$\beta \int_0^t q_0 \bar{y} \sigma_{WW} \lambda_1$	$\int_0^t q_0^2 \bar{y} \sigma_{WW} \lambda_1 \alpha_1$
$R^*(t)$	$\beta \int_0^t q_0 \bar{y} \sigma_{WR^*} \lambda_1$	$\int_0^t q_0^2 \bar{y} \sigma_{R^*R^*} \lambda_1 \alpha_1$

where $q_0(t)$ is the limit of the weight function $q(t)$; $\bar{y}(t)$ is the asymptotic proportion at risk at time t, $R^*(t) = R(t)/n(t)$; and $\sigma_{WW}(t)$, $\sigma_{R^*R^*}(t)$ and $\sigma_{WR^*}(t)$ are the asymptotic variances and covariances of the W's and R^*'s at risk at time t. These limiting functions depend on the transition hazard rates $\theta(t)$, $\psi_{S_0}(t)$, $\psi_{S_1}(t)$, $\lambda_{S_0}(t)$ and $\lambda_{S_1}(t, w)$ described

in § 1 and are derived in Appendix 2. The efficacy of a test is defined to be $\mu^2(t)/\sigma^2(t)$, and the asymptotic relative efficiency of one test relative to another is the ratio of their efficacies.

Two submodels of H_{ca} we will consider are:

Model 1: $\lambda_{S_1}^N(t, w) = 1 + \beta N^{-\frac{1}{2}} w + O(N^{-1})$,

Model 2: $\lambda_{S_1}^N(t, w) = \alpha_1/(b+t) + \beta N^{-\frac{1}{2}} w + O(N^{-1})$.

Model 1 is a special case of both a relative risk and an excess risk model, while Model 2 is an excess risk model.

Three versions of $T(q, Z)$ defined in § 2 are the Cox score test, given by $T(1, W)$ where $q_0 \equiv 1$, the generalized log rank test, given by $T(1, R^*(t))$ where $q_0 \equiv 1$, and the modified Kendall test, given by $T(n(t)/N, R^*(t))$ where $q_0(t) = \bar{y}(t)$. For convenience, T_{GL} will be used to denote the generalized log rank statistic and T_{K} will be used for the Kendall statistic. Also, as noted in § 2, there may be some advantage to considering time-weighted versions of the rank tests. Thus we define the time-weighted generalized log rank statistic $T_{t\text{-GL}}$ as $T(t, R^*(t))$ and the time-weighted Kendall statistic $T_{t\text{-K}}$ as $T(tn(t)/N, R^*(t))$. For $T_{t\text{-GL}}$, $q_0(t) = t$ and for $T_{t\text{-K}}$, $q_0(t) = t\bar{y}(t)$. The efficiency of the above tests can be improved by judicious choice of $q(t)$. Knowledge of $\mu(t)$ and $\sigma^2(t)$ as functions of the limiting weight function $q_0(t)$ can be used to find that weight function that maximizes the asymptotic efficacy. Using the Cauchy–Schwarz inequality one can show that under H_{ca} the optimal weight function when the label is W is $q_0(t) \propto 1/\alpha_1(t)$, and when the label is $R^*(t)$ it is $q_0(t) \propto \sigma_{WR^*}(t)/\alpha_1(t)$. The corresponding optimally weighted versions of T will be denoted by $T_{\text{opt,Cox}}$ and $T_{\text{opt,GL}}$, respectively. Note that the Cox test is optimal for the relative risk model where $\alpha_1(t) = 1$, but is suboptimal for the excess model. The function $\sigma_{WR^*}(t)$ is easily estimated at each failure time by the sample covariance between W and R^*. Unfortunately $\alpha_1(t)$ is unknown and not easily estimated, so $T_{\text{opt,Cox}}$ and $T_{\text{opt,GL}}$ are primarily of theoretical interest and are used here only for the sake of comparison. Let $T_{\text{cov-GL}}$ denote the test $T(\hat{\sigma}_{WR^*}(t), R^*(t))$. Among versions of $T(q, R^*(t))$ the $T_{\text{cov-GL}}$ test should be optimal under the relative risk model.

Table 2 gives the asymptotic relative efficiencies of the T_{GL}, T_{K}, $T_{t\text{-GL}}$, $T_{t\text{-K}}$ and $T_{\text{cov-GL}}$ tests relative to the Cox test under Model 1 for the cases of exponential and uniform waiting times. The time-weighted tests $T_{t\text{-GL}}$ and $T_{t\text{-K}}$ perform much better than their

Table 2. *Asymptotic relative efficiencies, evaluated at* $t = \infty$, *of the generalized log rank* T_{GL}, *Kendall* T_{K}, *their time-weighted versions* $T_{t\text{-GL}}$ *and* $T_{t\text{-K}}$, *and the covariance-weighted generalized log rank* $T_{\text{cov-GL}}$ *tests relative to the Cox test for a relative risk model*

Distribution of W	ψ_0	ψ_1	ARE (% of Cox test) T_{GL}	T_{K}	$T_{t\text{-GL}}$	$T_{t\text{-K}}$	$T_{\text{cov-GL}}$
Exp (2)	0	0	69·7	37·7	83·1	65·5	89·0
	0	1	66·8	31·9	99·4	57·3	99·4
	1	1	68·2	34·3	92·7	60·9	94·2
Unif (0, 1)	0	0	92·5	74·2	99·3	78·1	99·3
	0	1	86·6	73·4	96·3	79·3	96·3
	1	1	85·7	66·6	99·3	80·3	99·3

Transition rates described in § 1: $\lambda_{S_0}(t) = 0{\cdot}3$, $\psi_{S_0}(t) = \psi_0$, $\psi_{S_1}(t) = \psi_1$, $\lambda_{S_1}(t, w) = 1 + \beta N^{-\frac{1}{2}} w$.

unweighted counterparts. The $T_{\text{cov-GL}}$ test performs the best of the rank-based tests, and shows very good local power relative to the Cox test.

Model 2 is an excess risk model. The $T_{\text{opt,Cox}}$ test and $T_{\text{opt,GL}}$ test use the respective weight functions $b+t$ and $(b+t)\sigma_{WR^*}(t)$. This model would then favour a time-weighted test like $T_{t\text{-GL}}$ over T_{GL}. Figure 1(a) displays the asymptotic efficacies for the various test statistics as functions of time. The $T_{\text{opt,Cox}}$ and $T_{\text{opt,GL}}$ perform the best, of course, followed by $T_{t\text{-GL}}$, T_{Cox} and $T_{\text{cov-GL}}$. The T_{GL} and T_{K} tests perform very poorly. It is very important to note, however, that, had $\alpha_1(t) = ct$, then the optimal weight component $1/\alpha(t) = (ct)^{-1}$ would have made $T_{t\text{-GL}}$ a poor choice of test.

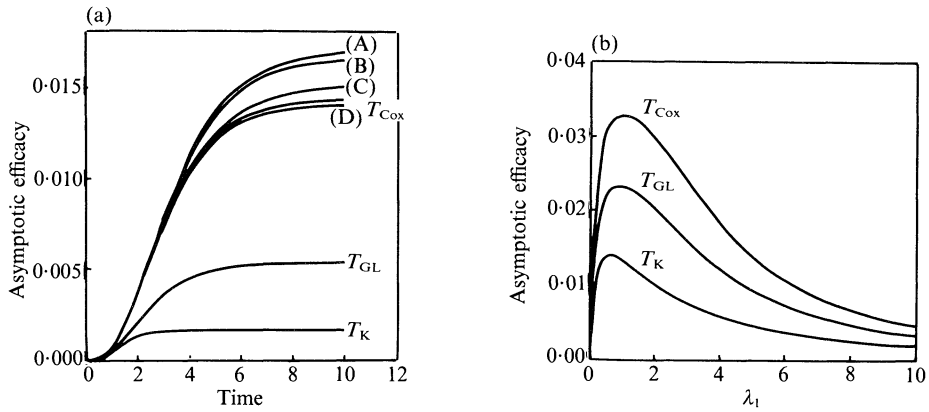

Fig. 1. Asymptotic efficacies of test statistics of Markov model. (a) Transition rates $\lambda_{S_0}(t) = 1$, $\theta(t) = 0{\cdot}5$, $\psi_{S_0}(t) = 0{\cdot}5$, $\psi_{S_1}(t) = 0{\cdot}2$, $\lambda_{S_1}(t, w) = 4/(1+t) + \beta N^{-\frac{1}{4}} w$; curves represent: (A) $T_{\text{opt,Cox}}$, (B) $T_{\text{opt,GL}}$, (C) $T_{t\text{-GL}}$, (D) $T_{\text{cov-GL}}$. (b) Transition rates $\lambda_{S_0}(t) = 0{\cdot}3$, $\psi_{S_0}(t) = 0{\cdot}5$, $\psi_{S_1}(t) = 0{\cdot}2$, $\theta(t) = 2$, $\lambda_{S_1}(t, w) = \lambda_1(1 + \beta N^{-\frac{1}{4}} w)$.

One interesting comparison between the Markov and semi-Markov settings is that the powers of the tests behave differently depending on the magnitude of the hazard. In the semi-Markov case the efficacy of a test increases as $\lambda_1(t)$ of the H_{ca} model increases (Jones & Crowley, 1990). However, as shown in Fig. 1(b), for the Markov case the efficacy of a test increases for a while as $\lambda_1(t)$ increases but then decreases as the failure hazard becomes so large that the proportion of subjects at risk in S_1 begins to decrease towards zero.

5. SMALL SAMPLE PROPERTIES

A Monte Carlo study was undertaken to determine the small sample rejection rates of the Markov versions of the Cox score test, generalized log rank T_{GL}, Kendall T_{K}, their time-weighted counterparts $T_{t\text{-GL}}$ and $T_{t\text{-K}}$ and the covariance-weighted generalized log rank $T_{\text{cov-GL}}$, as defined in § 4. For this study the S_0 to C transition hazard is assumed to be constant, that is $\psi_{S_0}(t) = \psi_0$. If $\psi_0 = 0$, no transition to C occurs from state S_0. The S_1 to C transition hazard is also assumed constant, that is $\psi_{S_1}(t, w) = \psi_1$. Two distributions for the waiting time are considered: exponential (2) and uniform (0, 1). The S_0 to D hazard rate is assumed constant: $\lambda_{S_0}(t) = 0{\cdot}3$. The S_1 to D failure hazard is modelled as one of

$$\lambda_{S_1}(t, w) = 1 + \beta w, \quad \lambda_{S_1}(t, w) = t + \beta w, \quad \lambda_{S_1}(t, w) = t(1 + \beta w),$$

representing constant, excess and relative risk hazards, respectively. In these simulations the range of (ψ_0, ψ_1) is $(0, 0)$, $(0, 1)$ and $(1, 1)$. At time $t = 0$, $N = 50$ individuals enter

state S_0, of whom N_1 will proceed to S_1 and N_2 to state D. There were 2000 simulations performed per experiment.

Table 3. *Observed power based on* 2000 *simulations per experiment*

					Percentage rejecting H_0 at 0·05 level					
$\lambda_{S_1}(t, W)$	ψ_0	ψ_1	Ave N_1	Ave N_2	T_{Cox}	T_{GL}	T_{K}	$T_{t\text{-GL}}$	$T_{t\text{-K}}$	$T_{\text{cov-GL}}$
				$W \sim$ Exponential (2)						
$1+3W$	0	0	43·5	43·5	75·9	67·0	56·3	68·4	66·0	73·3
	0	1	43·5	28·7	47·4	43·7	34·5	45·1	42·5	45·8
	1	1	30·3	19·1	34·4	29·4	25·1	30·4	30·3	32·2
$t+3W$	0	0	43·5	43·5	78·0	77·4	71·5	70·0	74·4	76·2
	0	1	43·5	26·4	54·9	57·8	50·5	52·5	54·5	54·0
	1	1	30·2	17·0	41·7	41·5	37·8	36·0	40·1	40·6
$t(1+3W)$	0	0	43·4	43·4	83·7	75·5	65·8	75·5	73·4	80·8
	0	1	43·5	25·2	54·2	50·0	39·9	51·8	47·4	53·2
	1	1	30·3	16·0	38·4	31·7	25·6	31·3	31·0	35·3
				$W \sim$ Uniform (0, 1)						
$1+3W$	0	0	43·2	43·2	62·8	59·5	54·1	61·2	55·8	60·8
	0	1	43·2	29·6	38·9	37·1	34·7	37·9	35·9	37·6
	1	1	28·0	18·6	31·5	29·6	26·7	30·1	28·9	29·8
$t+3W$	0	0	43·2	43·2	69·4	68·3	65·6	66·6	65·8	67·6
	0	1	43·2	28·1	46·2	47·4	44·9	44·8	43·9	44·8
	1	1	28·0	17·2	34·5	34·1	32·1	32·7	32·9	33·8
$t(1+3W)$	0	0	43·2	43·2	71·4	68·3	61·0	68·3	64·1	69·6
	0	1	43·1	26·9	43·9	40·7	36·2	40·8	36·7	41·3
	1	1	27·9	16·4	34·6	32·3	28·4	32·0	30·7	33·0

$N = 50$ individuals enter S_0, N_1 enter S_1, N_2 enter D from S_1. Transition rates described in § 1: $\lambda_{S_0}(t) = 0{\cdot}3$, $\psi_{S_0}(t) = \psi_0$, $\psi_{S_1}(t) = \psi_1$.

The rejection rates at the 0·05 level under H_0: $\beta = 0$ range from 4·10% to 6·80%. The results on power appear in Table 3 for $\beta = 3$. Powers decrease for all tests as the effective sample sizes decrease, due to increased censoring. For the fixed risk hazard $\lambda_{S_1}(t, w) = 1+3w$ and the relative risk hazard $\lambda_{S_1}(t, w) = t(1+3w)$, the Cox statistic is more powerful than the other tests, which is consistent with the optimality of the Cox score test for the relative risk model. Note, however, that the Cox test's superiority over the tests based on the ranked values of the waiting time W is greater when W is exponential rather than uniform, as one would expect. In all cases the covariance-weighted generalized log rank test is only slightly less powerful than the Cox test. The performance of the generalized log rank and its time-weighted version is similar for the relative risk model, whereas the generalized log rank test is better than the time-weighted version for the excess risk hazard $t+3w$. Given the discussion of § 4 on optimal weights, this is precisely what one would predict. All tests perform better than the Kendall, including the time-weighted Kendall. In these simulations the logit rank test was very similar to and only slightly less powerful than the generalized log rank, and was therefore not included.

6. Discussion

The medical researcher is frequently unsure whether the hazard λ_{S_1} of failure is a function of time since entering state S_0, that is $\lambda_{S_1}(t)$, or of time since entering S_1, that

is $\lambda_{S_1}(t-w)$, where t is the time since entering S_0. The former case is the Markov model whereas the latter case is the semi-Markov model. Since model misspecification can greatly influence study results, tests of whether the Markov model holds is an important part of the statistical analysis. In this paper we have adapted some previous work (Jones & Crowley, 1989) to the Markov model. This results in the general class of statistics $T(q, Z)$ given in § 2 appropriate for testing the Markov hypothesis. Special members of this class are the Cox partial likelihood score test, the generalized log rank test and the Kendall statistic as modified for the Markov setting. We also introduced time-weighted versions of the generalized log rank and Kendall statistics in § 4, and three test statistics based upon maximizing local efficiency with respect to a general hazard which includes both relative and excess risk models. Two of these tests, the so-called $T_{\text{opt,Cox}}$ and $T_{\text{opt,GL}}$, are primarily of theoretical interest since they depend on knowledge of the $\alpha_1(t)$ component of the hazard model H_{ca} given in § 4. The third of these tests, called $T_{\text{cov-GL}}$, is an adaptively weighted version of the generalized log rank test. The entire class of statistics given by $T(q, Z)$ are asymptotically normal and the asymptotic relative efficiencies under a contiguous sequence of a very general alternative hazards H_{ca} are given.

The choice between the unweighted and time-weighted generalized log-rank, T_{GL} and $T_{t\text{-GL}}$ respectively, depends on the true value of $\alpha_1(t)$ and the distribution of the waiting time W. The dependence on $\alpha_1(t)$ and W can be seen in the optimal choice of weight function as used by $T_{\text{opt,GL}}$. This dependence is also very evident in Tables 2 and 3 and Fig. 1(a). By use of an adaptive weight function the $T_{\text{cov-GL}}$ statistic comes closer to the optimal weight function. For this reason and because of its performance in the asymptotic and small sample simulated power studies, we recommend $T_{\text{cov-GL}}$ over T_{GL} and $T_{t\text{-GL}}$.

The Kendall statistics did not fare as well as their generalized log rank counterparts in any of our studies based on a relative risk hazard model in which $\alpha_1(t)=1$ or on an excess risk hazard model in which $\alpha_1(t)=b+t$. They were comparable to the T_{GL}-type tests when $\alpha_1(t)\propto t$. However, because the Kendall statistics use the number at risk as a weight function, they are very sensitive to the immigration and censoring rates. This is illustrated in the examples in § 3. Furthermore, the significant Kendall statistic obtained by Tsai (1990) for the Channing House data is questionable in light of the other test results given.

The Cox partial likelihood score test performed the best in our large and small sample studies. The $T_{\text{cov-GL}}$ test always maintained reasonable efficiency relative to the Cox test. When the distribution of waiting times contains outliers, the $T_{\text{cov-GL}}$ based on ranked waiting times should be less susceptible to distortion than the Cox test, which uses the actual waiting times. For this reason both the T_{Cox} and $T_{\text{cov-GL}}$ statistics are recommended as tests of the Markov model.

ACKNOWLEDGEMENT

This research was supported in part by a grant from the U.S. National Institutes of Health.

APPENDIX 1

U-statistic form of the Kendall statistic

In this appendix we shall write the U-statistic form of the Kendall rank correlation appropriate for the Markov model. Define T_j to be the jth individual's time to state D since entering S_0. Let

$c(u, v)$ be 1 if u is known to be larger than v, -1 if u is definitely smaller than v, and 0 if either $u = v$ or there is uncertainty which is larger. For example, if T_1 and T_2 are observed to be 3 and 4, respectively, and T_3 is censored at 2, then $c(T_1, T_2) = -1$ and $c(T_1, T_3) = 0$. A Kendall statistic appropriate for the Markov setting is

$$T_{\text{mod K}} = \sum_{i=1}^{N} \sum_{j=1}^{N} Y_j(T_i) c(T_i, T_j) c(W_i, W_j). \tag{A·1}$$

We need to introduce $Y_j(T_i)$ into the statistic so that the ith individual is compared to the jth only if the jth person is at risk when the ith fails. With some algebra one can show that, when there are no tied failure times, $T_{\text{mod K}} = 2T(n(t), R(t)/n(t))$, where $R(t)$ is defined in § 2. The statistic $T_{\text{mod K}}$ was introduced by M. P. Jones in his unpublished University of Washington Ph.D. thesis, and again by Tsai (1990). In the semi-Markov setting $Y_j(T_i)$ is unnecessary, in which case (A·1) is the Brown, Hollander & Korwar (1974) modification to the Kendall statistic for right-censored data.

Appendix 2

Calculation of terms used in asymptotic efficacy of $T(q, Z)$

The distribution $P_t(w)$ of the waiting time W in state S_1 at any time t will be derived in this appendix. From this the functions $\bar{y}$, σ_{WW}, σ_{WR^*} and $\sigma_{R^*R^*}$ needed in § 4 are calculated. Loosely speaking,

$$\begin{aligned} dP_t(w) &= \text{pr}\,(W = w \text{ at } t \,|\, \text{membership in } S_1 \text{ at time } t)\, dw \\ &= I(w < t) \frac{\exp\{-\int_0^w L_0(x)\,dx - \int_w^t L_1(x)\,dx\}\theta(w) \exp\{-\int_0^w \theta(x)\,dx\}}{\int_0^t \exp\{-\int_0^s L_0(x)\,dx - \int_s^t L_1(x)\,dx\}\theta(s) \exp\{-\int_0^s \theta(x)\,dx\}\,ds}, \end{aligned} \tag{A·2}$$

where $L_i(x) = \psi_{S_i}(x) + \lambda_{S_i}(x)$ for $i = 0, 1$. Note that the denominator of (A·2) is $\bar{y}(t) = \lim n(t)/N$ as $N \to \infty$. In Model 1, $\psi_{S_0}(t) = \psi_0$, $\psi_{S_1}(t) = \psi_1$, $\lambda_{S_0}(t) = \lambda_0$, $\lambda_{S_1}(t) = \lambda_1(1 + \beta N^{-\frac{1}{2}} W) \to \lambda_1$. If W is an exponential (θ) random variable, that is $\theta(t) = \theta$, then, by (A·2), $P_t(w)$ is a truncated exponential with density $I(w < t) a e^{-aw}(1 - e^{-at})^{-1}$, where $a = \psi_0 + \lambda_0 + \theta - \psi_1 - \lambda_1 \neq 0$. One can easily show that

$$\bar{y}(t) = \frac{\theta}{a} e^{-(\psi_1 + \lambda_1)t}(1 - e^{-at}), \tag{A·3}$$

$$\sigma_{WW}(t) = a^{-2} - t^2 e^{-at}(1 - e^{-at})^{-2}. \tag{A·4}$$

If $a = 0$, then $P_t(w)$ is uniform $(0, t)$ and analogues to (A·3) and (A·4) are easily found. If $P_0(w)$ is uniform $(0, b)$, then $P_t(w)$ is again a truncated exponential with parameter $a = \psi_0 + \lambda_0 - \psi_1 - \lambda_1$ but truncated at $\min(t, b)$. If W is uniform $(0, b)$ and $a = 0$, then $P_t(w)$ is uniform $\{0, \min(t, b)\}$.

It was shown by Jones & Crowley (1990) that the covariance at time t between the waiting times W and their ranks among those at risk at t is

$$\sigma_{WR^*}(t) = \int_{P_t^{-1}(0)}^{P_t^{-1}(1)} w F_t(w)\, dF_t(w) - \tfrac{1}{2} E_t(W),$$

where $F_t(w)$ is the distribution function of $P_t(w)$ and $E_t(W)$ its expectation. For the truncated exponential (a) where $a \neq 0$

$$\sigma_{WR^*}(t) = \frac{1 - 2aH(t)\, e^{-aH(t)} - e^{-2aH(t)}}{4a(1 - e^{-aH(t)})^2}, \tag{A·5}$$

where $H(t) = t$ when W is exponential and $H(t) = \min\{t, b\}$ when W is uniform $(0, b)$. Since W is assumed to be a continuous covariate, (A·18) of Lehmann (1975) implies $\sigma_{R^*R^*}(t) = \frac{1}{12}$. Calculation of (A·2) through (A·5) for Model 2 in which

$$\lambda_{S_1}(t) = \alpha_1/(b + t) + \beta N^{-\frac{1}{2}} W \to \alpha_1/(b + t)$$

is messier but straightforward and results in closed form expressions for integer values of α_1.

REFERENCES

BROWN, Jr., B. W., HOLLANDER, M. & KORWAR, R. M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies. In *Reliability and Biometry, Statistical Analysis of Lifelength*, Ed. F. Proschan and R. J. Serfling, pp. 327–54. Philadelphia: SIAM.

COX, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B **34**, 187–220.

HYDE, J. (1977). Testing survival under right censoring and left truncation. *Biometrika* **64**, 225–30.

JONES, M. P. (1991). Robust tests for survival data involving a single continuous covariate. *Scand. J. Statist.* **18**, 323–32.

JONES, M. P. & CROWLEY, J. (1989). A general class of nonparametric tests for survival analysis. *Biometrics* **45**, 157–70.

JONES, M. P. & CROWLEY, J. (1990). Asymptotic properties of a general class of nonparametric tests for survival analysis. *Ann. Statist.* **18**, 1203–20.

LEHMANN, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

O'BRIEN, P. C. (1978). A nonparametric test for association with censored data. *Biometrics* **34**, 243–50.

PRENTICE, R. L. & MAREK, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics* **35**, 861–7.

TSAI, W. Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika* **77**, 169–77.

[*Received July* 1990. *Revised July* 1991]