Biometrika (1990), 77, 4, *pp.* 853-64
Printed in Great Britain

Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment

BY DAVID M. ZUCKER AND EDWARD LAKATOS

Biostatistics Research Branch, National Heart, Lung, and Blood Institute, Bethesda, Maryland 20892, U.S.A.

SUMMARY

In certain long-term treatment trials, one expects some lag period before the treatment is fully effective. This paper presents two weighted log rank type statistics designed to have good efficiency over a wide range of lags: a maximin efficiency robust statistic and a simplified version of this statistic. Both statistics can be computed as easily as the log rank statistic. Asymptotic efficiency calculations, supported by small sample simulations, show that the two proposed statistics are substantially more efficient than the conventional log rank statistic in certain lag situations with comparatively little efficiency loss relative to the log rank statistic when no lag exists. We recommend the maximin statistic for situations where a lag is expected but cannot be specified precisely in advance.

Some key words: Clinical trial; Lag period; Relative efficiency; Weighted log rank statistic.

1. INTRODUCTION

It is well known that the log rank statistic is optimal for comparing survival curves under proportional hazards alternatives. Some investigators have introduced statistics in particular settings in which the proportional hazards assumption does not apply (Tarone & Ware, 1977; Harrington & Fleming, 1982; Fleming, Harrington & O'Sullivan, 1987; Mantel & Stablein, 1988).

The possibility of time lags in treatment effect has been noted by Halperin et al. (1968), Wu, Fisher & Demets (1980), Gail (1985) and Lakatos (1986, 1988). They discussed sample size calculation when such a lag was expected, assuming analysis would be by a standard method not accounting for lag. Here we discuss alternative methods which do account for lag.

Tarone & Ware (1977) describe a class of linear rank statistics in which the contribution of each event to the total statistic is given a specified weight. The member of the class is determined by the weighting method. The extreme members are the log rank statistic, with equal weighting, and Gehan's (1965) modified Wilcoxon statistic, with weighting proportional to the population at risk just before failure. Harrington & Fleming's (1982) G^{ρ} class behaves analogously; for this class, the extremes are the log rank statistic and Peto & Peto's (1972) modified Wilcoxon statistic. Although these classes might seem a good source for statistics which perform well under lag alternatives, such is not the case. Lag alternatives require down-weighting of early events, but within these classes, the log rank gives proportionately least weight to early events.

In the Women's Health Trial (Self et al., 1988), the designers expected a linear lag as described in § 2 below, and proposed a test statistic with corresponding linearly increasing weights. To our knowledge, the issue of time lags in analysis has not been discussed elsewhere in the literature.

We present and motivate by maximin arguments two weighted log rank type statistics which are designed to have good efficiency across a range of lag alternatives. Section 2 presents two prototypical time lag alternatives. Section 3 provides the setting and theoretical background. In § 4, we present the proposed statistics and discuss basic large sample efficiency considerations. Section 5 gives numerical calculations for a particular example, and § 6 a variety of small sample simulation results. Section 7 contains an overall discussion of the statistics and their properties. Our focus throughout is on one-sided alternatives.

2. Lag models

Let $\lambda_0(t)$ and $\lambda_1(t)$ denote the hazard functions for control and treatment, respectively. The proportional hazards model postulates that $\lambda_1(t) = \phi\lambda_0(t)$ for some constant $\phi < 1$. A general class of lag models may be described through the equation

$$\lambda_1(t) = [\phi l(t) + \{1 - l(t)\}]\lambda_0(t),$$

where $l(t)$ is a monotone function with $0 \leq l \leq 1$. The value of $l(t)$ represents the proportion of the treatment effect achieved by time t; for example, $l(t) = 1$ means that the treatment has reached its full effect.

In this paper, we focus attention on the following two prototypical lag functions, where I denotes an indicator function:

(a) Linear lag of length t^*:

$$l(t) = (t/t^*)I(t \leq t^*) + I(t > t^*).$$

The treatment effect increases linearly from 0 at time 0 to full effect at time t^*.

(b) Threshold lag of length t^*: $l(t) = I(t > t^*)$. The treatment has no detectable effect during the period $[0, t^*]$; afterwards, the treatment is fully effective.

The linear lag was described by Halperin et al. (1968). A prime example is cholesterol lowering therapy to prevent coronary events. Here, because treatment typically is initiated after 20 or more years of high cholesterol and associated plaque development, one expects the therapy to reduce risk gradually over time rather than immediately upon initiation. A linear lag was assumed in the sample size calculations for the Lipid Research Clinics Coronary Primary Prevention Trial; the trial results indicate the presence of at least some sort of lag phase (Lipid Research Clinics Program, 1979, 1984). The Women's Health Trial (Self et al., 1988) provides another example.

In the Physicians' Health Study (Physicians' Health Study Steering Committee, 1983), the planned analysis for testing the effect of beta-carotene on cancer incidence is one that is optimal under a threshold lag. Here, the investigators thought that treatment would not affect pre-existing tumours but would prevent new tumour development. Because of the time required for new tumours to become detectable, the investigators decided not to count cancers occurring during the first two years after randomization.

The threshold lag is the limiting form for families of S-shaped lag functions, e.g. logistic or probit lags.

The impact of a threshold lag on the efficiency of the log rank statistic is especially evident. When there is a threshold lag, the log rank statistic suffers because the early period of no treatment difference makes zero contribution to the expected value of the statistic but positive contribution to the variance.

The linear and threshold lag models cover a reasonable range of models for the lag in treatment effect. The precise lag function is usually unknown. Indeed, it is often difficult to specify even the duration of the lag.

3. Basic setting and theoretical background

We work in the setting of a clinical trial with two groups: control, group 0, and treatment, group 1. We assume a random censorship model. For simplicity, we assume that each group has the same sample size n. Associated with individual j in group i is a latent survival time T^0_{ij} and a latent censoring time V_{ij} which are independent random variables with distribution functions F_i and G_i, respectively. The $2n$ individuals in the study are mutually independent. Also, F_i is assumed to be absolutely continuous with density f_i and hazard $\lambda_i = f_i/(1-F_i)$. We adopt below the convention that all individuals enter at $t=0$; staggered entry can be handled by re-expression in terms of censoring.

The data consist of

$$T_{ij} = \min(T^0_{ij}, V_{ij}), \quad D_{ij} = I(T^0_{ij} \leqslant V_{ij}).$$

We define

$$\pi_i(t) = \text{pr}(T_{ij} \geqslant t) = \{1 - F_i(t)\}\{1 - G_i(t-)\}, \quad B_{ij}(t) = I(T_{ij} \leqslant t),$$

$$N_{ij}(t) = I(T_{ij} \leqslant t, D_{ij} = 1), \quad Y_i(t) = \sum_{j=1}^{n} I(T_{ij} \geqslant t), \quad N_i(t) = \sum_{j=1}^{n} N_{ij}(t).$$

The quantity $Y_i(t)$ is the number at risk in group i at time $t-$, and $N_i(t)$ is the number of events in group i up to time t. The total length of the trial, i.e. the follow-up period for the individual followed longest, is denoted by τ. We assume $F_i(\tau) < 1$.

For testing the null hypothesis of equal survivorship H_0: $F_0 = F_1$, a commonly considered class of statistics is the class of log rank type statistics having the stochastic integral form

$$T_W = n^{-1} \int_0^\tau W(s)\{Y_0(s)^{-1} + Y_1(s)^{-1}\}^{-1} \left\{\frac{dN_0(s)}{Y_0(s)} - \frac{dN_1(s)}{Y_1(s)}\right\}, \tag{1}$$

where W is a weight function for which $W(s)$ may depend on observations up to but not including time s. A common alternative way of writing (1) is

$$T_W = n^{-1} \textstyle\sum_k W(X_k)\left\{\delta_k - \dfrac{Y_0(X_k)}{Y_0(X_k) + Y_1(X_k)}\right\},$$

where X_k is the kth ordered failure time among both groups pooled together and δ_k is a $0-1$ indicator of whether the failure is in group 0, e.g. Tarone & Ware (1977). Taking $W \equiv 1$ gives the standard log rank statistic. The choice $W(s) = \{Y_0(s) + Y_1(s)\}^\gamma / n^\gamma$ with $\gamma \geqslant 0$ gives Tarone & Ware's (1977) family of tests, while the choice $W(s) = \hat{S}(s-)^\gamma$ with $\gamma \geqslant 0$ and $\hat{S}$ defined as the pooled-sample Kaplan-Meier survival function estimate gives Harrington & Fleming's (1982) family.

The asymptotic behaviour of statistics of the form (1) has been investigated extensively, for example by Aalen (1978) and Gill (1980), using the theory of martingales and stochastic integrals. This theory applies because, as indicated by Gill (1980, Corol. 3.1.1), the processes

$$M_{ij}(t)=N_{ij}(t)-\int_0^t \lambda_i(s)I(T_{ij}\geqslant s)\,ds$$

are square-integrable martingales with respect to the history $\{\mathscr{F}_t^n\}$ in which $\mathscr{F}_t^n$ is the completion of the σ-algebra generated by $B_{ij}(s)$, $D_{ij}B_{ij}(s)$ and $T_{ij}B_{ij}(s)$ for $s\leqslant t$, $i=0,1$ and $j=1,\dots,n$.

In particular, suppose that $\{W^{(n)}\}$ is a sequence of weight functions satisfying the following conditions.

Condition 1. $W^{(n)}$ is $\{\mathscr{F}_t^n\}$-predictable for each n.

Condition 2. $W^{(n)}\to W^\infty$ as $n\to\infty$ uniformly in probability on $[0,\tau]$ for some 'regular' deterministic function W^∞, where a function is called 'regular' if it is left-continuous, has right limits, and is of bounded variation.

Then, by Gill (1980, Corol. 4.3.1), under the null hypothesis T_W has an asymptotic mean-zero normal distribution in the sense that for

$$\sigma_W^2(H_0)=\int_0^\tau \{W^\infty(s)\}^2\{\pi_0(s)^{-1}+\pi_1(s)^{-1}\}^{-1}\lambda_0(s)\,ds$$

one has $\surd n\{T_W/\sigma_W(H_0)\}\to N(0,1)$ in distribution as $n\to\infty$. The variance $\sigma_W^2(H_0)$ may be estimated consistently by

$$\hat\sigma_W^2(H_0)=n^{-1}\int_0^\tau W^2(s)\{Y_0(s)^{-1}+Y_1(s)^{-1}\}^{-1}\frac{d(N_0+N_1)(s)}{(Y_0+Y_1)(s)}.$$

Thus, one may test H_0 by referring $Z_W=\surd n\{T_W/\hat\sigma_W(H_0)\}$ to the standard normal distribution.

If $W(s)=I(s>t)$, then we denote $U_t=Z_W$. The statistic U_t is obtained by computing the log rank statistic for the data in $(t,\tau]$ only. In particular, U_0 is the usual log rank statistic. The statistics U_t play an important role below.

The efficiency properties of log rank type statistics have been investigated by Gill (1980, § 5.2) within the framework of local asymptotics, i.e. in the limit as $n\to\infty$ and the treatment effect approaches zero at the rate $1/\surd n$. In particular, Gill (1980, eqn (5.2.15)) gives the Pitman efficacy. Specializing this, one finds that for 'regular' deterministic W, the Pitman efficacy of Z_W under a lag model with lag l is

$$e(W;l)=\left\{\int_0^\tau W(s)l(s)\psi(s)\,ds\right\}^2\bigg/\left\{\int_0^\tau W(s)^2\psi(s)\,ds\right\},$$

where $\psi(s)=\{\pi_0(s)^{-1}+\pi_1(s)^{-1}\}^{-1}\lambda_0(s)$, with π_1 evaluated under H_0. This is maximized for $W=l$; compare Gill (1980, Lemma 5.2.1). If W_1 and W_2 are two 'regular' deterministic functions, then the Pitman asymptotic relative efficiency of Z_{W_1} to Z_{W_2} when W_2 is optimal, and of Z_{W_2} to Z_{W_1} when W_1 is optimal, is given by

$$\rho^2(Z_{W_1},Z_{W_2})=\left(\int_0^\tau W_1W_2\psi\,ds\right)^2\bigg/\left(\int_0^\tau W_1^2\psi\,ds\right)\left(\int_0^\tau W_2^2\psi\,ds\right), \tag{2}$$

where the argument s inside the integrals has been suppressed. This is equal to the asymptotic correlation between Z_{W_1} and Z_{W_2} under H_0; compare Gastwirth (1985).

Suppose that $W_1, \ldots, W_{k+1}$ are deterministic weight functions and form a matrix R and a column vector c by defining $R_{pq} = \rho(Z_{W_p}, Z_{W_q})$ for $p, q = 1, \ldots, k$ and $c_q = \rho(Z_{W_{k+1}}, Z_{W_q})$ for $q = 1, \ldots, k$. Then if $a = (a_1, \ldots, a_k)^{\mathrm{T}}$ is a vector of nonnegative constants, the Pitman asymptotic relative efficiency of the statistic $V = a_0 Z_{W_0} + \ldots + a_k Z_{W_k}$ relative to $Z_{W_{k+1}}$ when W_{k+1} is optimal is equal to the square of the asymptotic null correlation between V and $Z_{W_{k+1}}$, which is given by

$$\rho^2(V, Z_{W_{k+1}}) = (c^{\mathrm{T}} a)^2 / (a^{\mathrm{T}} R a). \tag{3}$$

Suppose that $W_1^{(n)}$ and $W_2^{(n)}$ are data-dependent weight functions which satisfy Conditions 1 and 2 for 'regular' deterministic functions W_1^{∞} and W_2^{∞}, respectively. Then (2) with $W_h^{(n)}$ replaced by W_h^{∞} on the right-hand side gives the Pitman asymptotic relative efficiency of W_1 when W_2^{∞} is optimal, and of W_2 when W_1^{∞} is optimal. An analogous statement holds for (3).

As a special case of (2), if $t_1 < t_2$, then

$$\rho^2(U_{t_1}, U_{t_2}) = \frac{\Psi(\tau) - \Psi(t_2)}{\Psi(\tau) - \Psi(t_1)}, \tag{4}$$

where $\Psi(t) = \int \psi(s)\, ds$, with the integral over the range $(0, t)$. This special case will be used heavily in §§ 4 and 5.

4. Proposed test statistics

If l is known, then $W = l$ is optimal. For l unknown, we have investigated two approaches. One is to estimate the optimal W based on the data. This may be done validly provided that Conditions 1 and 2 are satisfied; in particular, $W(s)$ must not involve data beyond the interval $(0, s]$. However, a preliminary investigation indicated that for trials of realistic sample size, because of difficulties in achieving reasonable precision in estimating the optimal weights, this approach would be unlikely to be fruitful.

The other approach is to specify a reasonable range of lags and use some linear combination of the corresponding optimal Z_W's as the test statistic. The idea is that such a statistic could be hoped to have reasonable efficiency relative to the optimal Z_W across the range of alternatives. This is the approach taken in this paper.

In particular, we suppose that the set of plausible models for the lag function consists of the class $\mathscr{L}(t^{**})$ of functions which are monotone nondecreasing on $[0, t^{**}]$ and equal to one on $(t^{**}, \tau]$. In other words, the set of plausible lag models ranges from no lag to a threshold lag of length t^{**} for some $t^{**} < \tau$ and includes all intermediate possibilities.

As our proposed test statistics, we consider (a) the maximin efficiency robust test for the class $\mathscr{L}(t^{**})$, to be denoted by V^* (Gastwirth, 1966, 1985), and (b) an approximate version of the maximin test, given by $V_0 = (U_0 + U_{t^{**}})$. The maximin statistic is formulated as follows. For any log rank type statistic V, we define the minimum asymptotic relative efficiency of V as the minimum of $\rho^2(V, Z_l)$ over all $l \in \mathscr{L}(t^{**})$. The maximin statistic V^* is that statistic, among all V, for which the minimum asymptotic relative efficiency is as high as possible. In other words, among all V, the statistic V^* minimizes the worst possible efficiency loss, over all $l \in \mathscr{L}(t^{**})$, associated with using V instead of the optimal statistic Z_l.

As proved in Appendix 2, the maximin efficiency robust test V^* is given by Z_{W^*} with

$$W^*(s) = \left\{1 - \frac{\Psi(s)}{\Psi(\tau)}\right\}^{-\frac{1}{2}} I(s \leq t^{**}) + 2\left\{1 - \frac{\Psi(t^{**})}{\Psi(\tau)}\right\}^{-\frac{1}{2}} I(s > t^{**}). \tag{5}$$

As discussed in Appendix 2, V^* may be viewed as the limit as $k \to \infty$ of the linear combination statistic

$$V_k = U_0 + U_{t^{**}} + (1-\rho^{1/2^k}) \sum_{j=1}^{2^k-1} U_{t_{kj}}, \tag{6}$$

where $\rho^2 = \rho^2(U_0, U_{t^{**}}) = \{\Psi(\tau) - \Psi(t^{**})\}/\Psi(\tau)$ and

$$t_{kj} = \Psi^{-1}\{\Psi(\tau)(1-\rho^{j/2^{k-1}})\} \quad (j = 1, \ldots, 2^k - 1). \tag{7}$$

An important property of V^* is that $\rho(V^*, U_t)$ is constant over $t \in [0, t^{**}]$.

The results in Appendix 1 imply that the minimum of $\rho(V_0, Z_l)$ over all $l \in \mathscr{L}(t^{**})$ occurs for the threshold lag $l(s) = I(s > t_{11})$ with t_{11} given by (7). In addition, the minimum of $\rho(V^*, Z_l)$ over all $l \in \mathscr{L}(t^{**})$ occurs for the threshold lags $l(s) = I(s > t)$ for $t \leq t^{**}$. The corresponding minimum asymptotic relative efficiencies are

$$\rho^2(V_0, U_{t_{11}}) = 2/(1+\rho^{-1}), \quad \rho^2(V^*, U_t) = 2/(2 - \log \rho). \tag{8}$$

These expressions allow one to quantify the worst possible performance of V_0 and V^*, in terms of asymptotic efficiency, over all lags in the class $\mathscr{L}(t^{**})$.

A sense of how the log rank statistic U_0, the statistic U_t, and the statistics V^* and V_0 behave may be gained by looking at how they weight the different parts of the study. As compared with the log rank statistic, the statistic U_t places zero weight on $[0, t]$ and full weight on $(t, \tau]$, whereas the statistics V^* and V_0 place partial weight on $[0, t^{**}]$ and full weight on $(t^{**}, \tau]$. This reveals another advantage of using the proposed statistics as compared with the Physicians' Health Study approach of using U_t for some t: the proposed statistics provide some protection against concluding that treatment is beneficial when there is actually an adverse effect during the hypothesized lag phase; the statistic U_t provides no such protection. This issue is discussed further in § 7.

Implementing V^* and V_0 is straightforward. For V_0, one calculates U_0, $U_{t^{**}}$, and

$$\hat{\rho} = \hat{\sigma}\{H_0;\ W(s) = I(s > t^{**})\}/\hat{\sigma}\{H_0;\ W(s) = 1\}.$$

The z-value for V_0 is then given by $(U_0 + U_{t^{**}})/\{2(1+\hat{\rho})\}^{1/2}$. To implement V^*, one estimates Ψ, based on pre-trial projections or based on the trial data themselves using the formula

$$\hat{\Psi}(t) = n^{-1} \int_0^t \{Y_0(s)^{-1} + Y_1(s)^{-1}\}^{-1} \frac{d(N_0 + N_1)(s)}{(Y_0 + Y_1)(s)} \tag{9}$$

or a smoothed version thereof, and substitutes the estimate into the formula (5) for W^*. Note that $\hat{\Psi}(t)$ is obtained in the same way as $\hat{\sigma}^2(H_0)$ for the log rank test, but counting only the data in the interval $[0, t]$. Gill (1980, Lemma 4.3.1) shows that $\hat{\Psi} \to \Psi$ uniformly in probability on $[0, \tau]$ as $n \to \infty$; it hence can be shown that substituting $\hat{\Psi}$ for Ψ does not affect the asymptotic distribution of V^*, even though the estimated W^* is not predictable.

Unlike many common log rank type statistics, V^* and V_0 are not linear rank statistics, because the weight given to each event depends on the time of the event and not just on its rank among all events. However, one may modify V^* and V_0 to obtain linear rank statistics by using (9) in V^* and replacing t^{**} with $\hat{F}^{-1}(p^{**})$, where $\hat{F}$ is the pooled-sample Kaplan-Meier survival time distribution function estimate and $p^{**} \in (0, 1)$ is an a priori estimate of the maximum plausible lag length on the scale of cumulative event rate.

5. Numerical calculations of asymptotic efficiency

To give a numerical flavour of the efficiency properties of the statistics U_t, V_0 and V^*, calculations are presented for the following example. A study is conducted to compare treatment to control with respect to survival. The total trial period is τ years. There is uniform censoring over the interval $[\tau_1, \tau]$ in both groups, with $\tau_1 = 0{\cdot}7\tau$. In the control group, survival is exponentially distributed with a τ-year mortality rate of $\theta = 50\%$. For this example, the function ψ is given by

$$\psi(t) = \tfrac{1}{2}\lambda\, e^{-\lambda t}[1 - \{(t-\tau_1)/(\tau-\tau_1)\}I(t > \tau_1)],$$

where $\lambda = -\tau^{-1}\log(1-\theta)$ is the control hazard rate.

Table 1 shows the asymptotic relative efficiency under each lag model across a range of values for the true lag t^* for (i) various U_t statistics, (ii) the statistic V_0, and (iii) the statistic V^*. These results were calculated using (2), (3) and (4), with numerical integration where necessary. Table 2 shows the minimum efficiencies given by (8) for $\theta = 0{\cdot}50$, as in Table 1, and also for $\theta = 0{\cdot}20$. The results are fairly insensitive to θ.

Under the threshold lag model, the asymptotic relative efficiency of the log rank statistic drops sharply as the lag t^* increases. By contrast, for values of t^{**} ranging up to $0{\cdot}5\tau$, the statistics V_0 and V^* provide substantial gains in efficiency relative to the log rank

Table 1. *Asymptotic relative efficiency of the U_t, V_0 and V^* statistics*

		Threshold lag						Linear lag				
		Length of lag (t^*/τ)						Length of lag (t^*/τ)				
	t^{**}/τ	0·0	0·1	0·2	0·3	0·4	0·5	0·1	0·2	0·3	0·4	0·5
ARE	0·0	1·000	0·849	0·709	0·577	0·455	0·341	0·950	0·902	0·859	0·819	0·783
of	0·1	0·849	1·000	0·834	0·680	0·536	0·401	0·945	0·972	0·948	0·915	0·880
U_t	0·2	0·709	0·834	1·000	0·815	0·642	0·481	0·789	0.883	0·935	0·937	0·920
	0·3	0·577	0·680	0·815	1·000	0·788	0·590	0·642	0·720	0·812	0·876	0·895
	0·4	0·455	0·536	0·642	0·788	1·000	0·749	0·506	0·567	0·640	0·729	0·798
	0·5	0·341	0·401	0·481	0·590	0·749	1·000	0·379	0·425	0·479	0·546	0·629
ARE	0·1	0·961	0·961	0·802	0·653	0·515	0·385	0·986	0·974	0·940	0·901	0·865
of	0·2	0·921	0·914	0·921	0·750	0·591	0·443	0·942	0·969	0·973	0·952	0·924
V_0	0·3	0·880	0·866	0·865	0·880	0·693	0·519	0·896	0·919	0·949	0·963	0·952
	0·4	0·837	0·816	0·806	0·810	0·837	0·627	0·849	0·866	0·890	0·923	0·944
	0·5	0·792	0·763	0·744	0·737	0·749	0·792	0·798	0·810	0·827	0·853	0·889
ARE	0·1	0·961	0·961	0·802	0·653	0·515	0·385	0·987	0·976	0·940	0·902	0·866
of	0·2	0·921	0·921	0·921	0·750	0·591	0·443	0·946	0·974	0·977	0·955	0·926
V^*	0·3	0·879	0.879	0·879	0·879	0·693	0·519	0·903	0·930	0·960	0·972	0·960
	0·4	0·835	0·835	0·835	0·835	0·835	0·626	0·858	0·884	0·913	0·945	0·963
	0·5	0·788	0·788	0·788	0·788	0·788	0·788	0·810	0·834	0·861	0·891	0·927

Exponential control group survival with τ-year mortality of $\theta = \frac{1}{2}$.
Note that U_0 is the usual log rank test.

Table 2. *Worst asymptotic relative efficiency of V_0 and V^* statistics over the class $\mathscr{L}(t^{**})$*

		Ratio t^{**}/τ								
Statistic	θ	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9
V_0	0·20	0·966	0·927	0·883	0·831	0·768	0·687	0·574	0·421	0·234
	0·50	0·959	0·914	0·864	0·806	0·737	0·652	0·537	0·386	0·209
V^*	0·20	0·967	0·932	0·895	0·854	0·809	0·756	0·687	0·602	0·497
	0·50	0·961	0·921	0·879	0·835	0·788	0·734	0·666	0·583	0·482

statistic for various lags t^* at comparatively little cost in terms of diminished efficiency relative to the log rank statistic when no lag exists. For example, the proposed statistics with $t^{**} = 0{\cdot}2\tau$ are 30% more efficient than the log rank statistic when the true lag $t^* = 0{\cdot}2\tau$ but only 8% less efficient than the log rank statistic when there is no lag.

The asymptotic relative efficiency of the threshold statistic U_t is highly sensitive to the choice of the cutpoint t. Thus, if one attempts to guess t^* itself and uses the corresponding U_t statistic, one can be severely penalized if the guess is wrong. By using V_0 or V^* instead, and choosing a t^{**} that exceeds all plausible t^*, this penalty may be circumvented. Avoiding a t^{**} that might underestimate the true lag is important because the asymptotic relative efficiency of each of the proposed statistics decreases rapidly as t^* increases beyond t^{**}, though not as rapidly as that of the log rank statistic.

Under the linear lag model, the statistics V_0 and V^* provide an appreciable increase in efficiency compared to the log rank test in the presence of a lag, although the increase is not as large as under the threshold model. The loss in efficiency associated with using V_0 or V^* when there is no lag is roughly comparable to the gain in efficiency associated with using V_0 or V^* when there is a linear lag of length t^{**}. As t^* increases beyond t^{**}, the asymptotic relative efficiency of each of the proposed statistics decreases; however, the decrease is gradual. For $t^* > t^{**}$, both V_0 and V^* have uniformly greater efficiency than the log rank statistic.

Under both lag models, the difference in asymptotic relative efficiency between V_0 and V^* shown in Table 1 is fairly small, especially for $t^{**} \leqslant 0{\cdot}3\tau$. However, as illustrated by Table 2, the difference between the two statistics becomes more pronounced for $t^{**} > 0{\cdot}5\tau$.

6. SMALL SAMPLE SIMULATION RESULTS

The asymptotic relative efficiencies in § 5 describe the behaviour of the log rank, V_0, and V^* statistics in the limit as the sample size n approaches infinity and the treatment effect approaches zero at the rate of $1/\surd n$. We have done simulations to investigate how well these results reflect small sample behaviour. All the simulations were run within the setting of the example in § 5 and assuming that when the full effect of treatment is achieved, the treatment hazard rate is 40% lower than the control rate.

The simulations examined the behaviour of the three statistics for $t^{**} = 0{\cdot}1\tau$, $0{\cdot}2\tau$ and $0{\cdot}4\tau$. For each value of t^{**}, simulations were run for no lag, a linear lag with $t^* = t^{**}$, and a threshold lag with $t^* = t^{**}$. Using Lakatos's (1988) procedure, the total trial size was set so that the asymptotically locally optimal log rank type test, i.e. the statistic Z_l with l equal to the true lag, would have approximately 80% power, with one-sided testing at the $\alpha = 0{\cdot}025$ level.

Powers were computed based on 5000 simulations, yielding standard errors of about 0·007. The test based on V^* was implemented using the estimate (9). The simulated powers are approximate because the critical values used were based on the asymptotic theory rather than on the exact finite sample null distribution of the test statistics. The simulations were carried out using the efficient method of K. K. G. Lan and E. Lakatos's, given in an unpublished report, with the same simulated data used for all three statistics. Also, a 'predicted power' for each statistic T was computed based on asymptotic relative efficiency results in § 5 using $\Phi\{-z_\alpha + (z_\alpha + z_\beta)\rho(T, Z_l)\}$, where $z_\alpha = 1{\cdot}96$ and $z_\beta = 0{\cdot}84$, and Φ is the standard normal distribution function.

Table 3 summarizes the results. There is excellent agreement between the 'predicted power' and the simulated power.

Table 3. *A comparison of simulation-based estimated power with power predicted based on asymptotic relative efficiency results*

t^{**}	Sample	Lag	Log rank		V_0		V^*	
value	size	model	Pred.	Sim.	Pred.	Sim.	Pred.	Sim.
0·1τ	314	No lag	0·800	0·792	0·784	0·776	0·784	0·775
0·1τ	350	Linear	0·776	0·771	0·794	0·792	0·794	0·792
0·1τ	372	Threshold	0·733	0·717	0·784	0·782	0·784	0·782
0·2τ	314	No lag	0·800	0·802	0·767	0·771	0·767	0·771
0·2τ	394	Linear	0·750	0·750	0·788	0·789	0·789	0·792
0·2τ	450	Threshold	0·655	0·626	0·767	0·752	0·767	0·750
0·4τ	314	No lag	0·800	0·803	0·727	0·720	0·726	0·720
0·4τ	512	Linear	0·699	0·685	0·768	0·758	0·775	0·768
0·4τ	708	Threshold	0·472	0·423	0·727	0·720	0·726	0·710

In addition, we carried out simulations under the null hypothesis of no treatment effect, for a trial with a total sample size of 100, in order to check the Type I error level of the proposed statistics. Based on 10 000 simulations, the following Type I error level estimates were obtained for a test with a nominal one-sided level of $\alpha = 0{\cdot}025$: 0·0257 for the log rank statistic, 0·0257 for the V_0 statistic, and 0·0256 for the V^* statistic.

7. Discussion

The popularity of the log rank test is warranted because (i) no modelling assumptions are needed regarding the form of the survival distributions, and (ii) under proportional hazards, the log rank statistic is optimal among the class of linear rank statistics. However, when there is a lag in the treatment effect, the proportional hazards assumption is violated, making the usual log rank test inefficient and a weighted version, which still avoids modelling, more suitable. Although several authors have described various classes of weighted log rank and related statistics, the problem of choosing weights for a lag situation has not been systematically investigated previously.

Several approaches are possible. If one thinks the lag function is equal to some function l, then one might consider the statistics Z_l, as given by Self et al. (1988). When the true lag is in fact l, then Z_l provides the greatest efficiency among all weighted log rank type statistics. However, because detailed a priori information about the lag is rarely available, Z_l may be a poor choice.

Another approach, taken in the Physicians' Health Study, is to give positive weight only to the portion of the trial during which one feels fairly certain that all or most of the full treatment effect will be present. This approach has serious drawbacks. On the one hand, an early adverse effect may be overlooked, leading to the conclusion that treatment is beneficial when in fact it is not. On the other hand, there can be a severe loss of efficiency due to misestimation of the lag length or to the discarding of a segment of data in which a partial treatment effect exists.

We have introduced statistics V^* and V_0 to provide good efficiency for a wide range of lags and to avoid severe efficiency loss. Moreover, because they include all events, there is less likelihood of concluding that treatment is beneficial when there is an early adverse effect. Admittedly, because they downweight early events, there is some potential for such misdirection. However, there is an unavoidable trade-off. We submit that the statistics V^* and V_0 strike a better balance between enhancing efficiency and avoiding misdirection than does the approach of completely ignoring the early data.

When there is little a priori knowledge regarding the lag and a substantial possibility that the lag is close in form to a threshold lag function, we recommend the statistics V^* and V_0. In terms of efficiency relative to the log rank statistic, they are greatly superior under a threshold lag and substantially superior under a linear lag. Moreover, these gains are realized with comparatively moderate efficiency loss when there is no lag. Some may prefer V_0 because it is simpler and comparable to V^* in efficiency unless the range of possible lags is extremely broad. However, calculating V^* with Ψ estimated by (9) is straightforward and we therefore recommend this procedure.

When there are specific data or medical considerations bearing on the nature of the lag function, clearly these should be taken into account in planning the statistical analysis. If the class of plausible lag functions differs from the class $\mathscr{L}(t^{**})$, the statistics V_0 and V^* may not be appropriate. However, in this case, the ideas of this paper can be used to construct and describe properties of a log rank type statistic adapted to the specified class of plausible lag functions.

APPENDIX 1

Least favourable lag functions

THEOREM 1. *For any deterministic 'regular' weight function W, the minimum of $\rho^2(Z_W, Z_l)$ over all $l \in \mathscr{L}(t^{**})$ occurs for a threshold lag function.*

Proof. Let $W \geqslant 0$ be given. To show that $\rho^2(Z_W, Z_l) \geqslant \min_t \rho^2(Z_W, U_t)$ for any $l \in \mathscr{L}(t^{**})$, it suffices to consider a step function l, because any l can be approximated by a step function. Suppose then that

$$l(t) = b_0 I(t > t_0) + \ldots + b_{m+1} I(t > t_{m+1}),$$

where $0 = t_0 < t_1 < \ldots < t_m < t_{m+1} = t^{**}$ and $b_j \geqslant 0$, with $b_0 + \ldots + b_{m+1} = 1$. In this case, Z_l is asymptotically equivalent to

$$(\beta_0 U_{t_0} + \ldots + \beta_{m+1} U_{t_{m+1}}) \Bigg/ \left\{ \sum_{j=0}^{m+1} \sum_{p=0}^{m+1} \beta_k \beta_p \rho(U_{t_j}, U_{t_p}) \right\}^{\frac{1}{2}},$$

where $\beta_j = b_j \{\Psi(\tau) - \Psi(t_j)\}^{1/2}$. Hence, by (3), $\rho^2(Z_W, Z_l)$ is equal to

$$\left\{ \sum_{j=0}^{m+1} \beta_j \rho(U_{t_j}, Z_W) \right\}^2 \Bigg/ \sum_{j=0}^{m+1} \sum_{p=0}^{m+1} \beta_j \beta_p \rho(U_{t_j}, U_{t_p}).$$

With $B = (\beta_0 + \ldots + \beta_{m+1})^2$, the denominator is bounded above by B and the numerator is bounded below by $B \min_t \rho^2(Z_W, U_t)$. The result follows. For 'regular' W, $\rho^2(Z_W, U_t)$ is continuous in t, so that the minimum is attained. □

APPENDIX 2

Statistics V_k and V^*

Gastwirth's (1985) recipe for generating 'candidates' for the maximin efficiency robust test over $\mathscr{L}(t^{**})$ leads to the following algorithm.

(a) Find all values of $s \leqslant t^{**}$, $s_1, \ldots, s_p$, say, which minimize $\rho(V_k, U_s)$.

(b) Define V_{k+1} as follows, with a's, b's and c's chosen so that V_{k+1} is equally correlated with U_0, $U_{t^{**}}$, the $U_{t_{kj}}$, and the U_{s_q}:

$$V_{k+1} = a_0 U_0 + a_1 U_{t^{**}} + \sum_{j=1}^{2^k - 1} b_j U_{t_{kj}} + \sum_{q=1}^{p} c_q U_{s_q}.$$

THEOREM 2. (a) *The Gastwirth recipe starting from V_0 leads to V_k as defined in* (6). *Moreover, $\rho^2(V_k, U_s)$ is maximized for $s = t_{kj}$ $(j=0, \ldots, 2^k)$ and minimized for $s = t_{k+1,j}$ with j odd, with corresponding extreme values*

$$\max(k) = \frac{1+\rho^{1/2^k}}{2+(2^k-1)(1-\rho^{1/2^k})}, \quad \min(k) = \left\{\frac{4\rho^{1/2^k}}{(1+\rho^{1/2^k})^2}\right\} \max(k);$$

(b) *V_k is the maximin test for $\mathscr{L}_k = \{l\colon l(s) = I(s > t_{kj}), \text{ some } j\}$;*
(c) *V^* is the maximin test for $\mathscr{L}(t^{**})$.*

Proof. An outline proof is given; details are available from the authors. We assume that $\psi \equiv 1$; the case of general ψ can be reduced to this case by the time transformation $t' = \Psi(t)$. For $t_1 \leq t_2 \leq t_3$, the following hold, the first for $\psi \equiv 1$ and the second for any ψ:

(i) $\rho^2(U_{t_1}, U_{t_2}) = (\tau - t_2)/(\tau - t_1)$,
(ii) $\rho(U_{t_1}, U_{t_3}) = \rho(U_{t_1}, U_{t_2})\rho(U_{t_2}, U_{t_3})$.

For $\psi \equiv 1$ we have

$$\rho^2 = \rho^2(U_0, U_{t^{**}}) = (\tau - t^{**})/\tau, \quad t_{kj} = (1-\rho^{j/2^{k-1}})\tau \quad (j = 0, \ldots, 2^k).$$

Note that $t_{k0} = 0$ and $t_{k2^k} = t^{**}$.

We prove (a) by induction. For $k = 0$, $V_0 = U_0 + U_{t^{**}}$ is given. Now note that

$$\begin{aligned} \rho(V_0, U_s) &= \{2(1+\rho)\}^{-\frac{1}{2}}\{\rho(U_0, U_s) + \rho(U_{t^{**}}, U_s)\} \\ &= \{2(1+\rho)\}^{-\frac{1}{2}}[\{(\tau - s)/\tau\}^{\frac{1}{2}} + \{(\tau - t^{**})/(\tau - s)\}^{\frac{1}{2}}]. \end{aligned} \tag{A1}$$

A simple calculus argument shows that this is maximized for $s = 0$ and $s = t^{**}$ and minimized for $s = t_{11}$; max (0) and min (0) are easily verified.

Assume now that $k = m - 1$; we verify the result for $k = m$. By the result for the extrema of $\rho(V_{m-1}, U_s)$, the s_q involved in V_m are the t_{mj} for j odd. That is, $V_m = \Sigma\, \omega_j U_{t_{mj}}$ for some vector ω, with j ranging from 0 to $2m$. By (i), the correlation matrix $R = R(m)$ of the $U_{t_{mj}}$ satisfies $R_{j_1 j_2} = \rho^{|j_1 - j_2|/2^m}$. By (3), $\rho(V_m, U_{t_{mj}}) = (R\omega)_j/(\omega^{\mathrm{T}} R\omega)^{\frac{1}{2}}$. The condition that all the $\rho(V_m, U_{t_{mj}})$ be equal thus requires that $R\omega = \eta e$, where $e = (1 \ldots 1)^{\mathrm{T}}$ and η is any constant. Straightforward algebra shows that a solution is given by $\omega_0 = \omega_{2^m} = 1$ and $\omega_j = (1 - \rho^{1/2^m})$ for $1 \leq j \leq 2^m - 1$. This verifies the expression for V_m.

For $s \in [t_{mp}, t_{m,p+1}]$, from (3), (ii), the formula for R, and the geometric series formula, we find

$$(\omega^{\mathrm{T}} R\omega)^{\frac{1}{2}} \rho(V_m, U_s) = \rho(U_{t_{mp}}, U_s) + \rho(U_{t_{m,p+1}}, U_s).$$

Arguing as with (A1) verifies the extrema of $\rho(V_m, U_s)$.

Regarding (b), Gastwirth (1966, p. 936) shows that the maximin test over $\mathscr{L}_k$ is given by $\Sigma\, c_j U_{t_{kj}}$, where the vector c is the solution to the following quadratic program: minimize $c^{\mathrm{T}} R(k) c$, subject to $R(k)c \geq 1$ and $c \geq 0$. A Kuhn-Tucker argument (Luenberger, 1984, § 10.8) shows that the solution is $c = R(k)^{-1} e$, which gives a statistic equivalent to V_k.

Regarding (c), the main result, it is clear that for any k we have

$$e^* = \sup_W \inf_{l \in \mathscr{L}(t^{**})} \rho^2(Z_W, Z_l) \leq \sup_W \inf_{l \in \mathscr{L}_k} \rho^2(Z_W, Z_l).$$

Because V_k is the maximin test for $\mathscr{L}_k$, the right-hand side equals min (k). By l'Hôpital's rule, $(1 - \rho^x)/x \to -\log \rho$ as $x \to 0$. Using this one finds that $\min(k) \to 2/(2 - \log \rho)$ as $k \to \infty$. Thus, $e^* \leq 2/(2 - \log \rho)$. Under the assumption $\psi \equiv 1$, formula (5) for W^* becomes

$$W^*(s) = (1 - s/\tau)^{-1/2} I(s \leq t^{**}) + 2(1 - t^{**}/\tau)^{-1/2} I(s > t^{**}). \tag{A2}$$

Using this and evaluating the integrals in (2) gives $\rho^2(V^*, U_t) = 2/(2 - \log \rho)$ for all $t \in [0, t^{**}]$. Thus, from Theorem 1, $e^* = 2/(2 - \log \rho)$ and V^* is the maximin test.

The expression (5) for W^* may be derived in two different ways. One is to treat the weight function for V_k as a Riemann sum and find the integral to which it converges. Alternatively, noting that $\max(k) - \min(k) \to 0$ as $k \to \infty$, one may anticipate that $\rho(V^*, U_t)$ will be constant over $t \in [0, t^{**}]$ and solve the differential equation $(d/dt)\rho(V^*, U_t) = 0$. For $\psi \equiv 1$, this reduces, using (2), to $\mathcal{W}'(t) = (\mathcal{W}(1) - \mathcal{W}(t))/\{2(1-t)\}$ for $t \in [0, t^{**}]$, where $\mathcal{W}(t)$ is the integral of W over $[0, t]$. This is easily solved to yield (A2).

REFERENCES

AALEN, O. O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6**, 701-26.

FLEMING, T. R., HARRINGTON, D. P. & O'SULLIVAN, M. (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. *J. Am. Statist. Assoc.* **82**, 312-20.

GAIL, M. H. (1985). Applicability of sample size calculations based on a comparison of proportions for use with the log rank test. *Controlled Clin. Trials* **6**, 112-9.

GASTWIRTH, J. L. (1966). On robust procedures. *J. Am. Statist. Assoc.* **61**, 929-48.

GASTWIRTH, J. L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *J. Am. Statist. Assoc.* **80**, 380-4.

GEHAN, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 202-23.

GILL, R. D. (1980). *Censoring and Stochastic Integrals*, Mathematical Centre Tract 124. Amsterdam: Mathematisch Centrum.

HALPERIN, M., ROGOT, E., GURIAN, J. & EDERER, F. (1968). Sample sizes for medical trials with special reference to long-term therapy. *J. Chronic Dis.* **21**, 13-24.

HARRINGTON, D. P. & FLEMING, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553-66.

LAKATOS, E. (1986). Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Controlled Clin. Trials* **7**, 189-99.

LAKATOS, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics* **44**, 229-41.

LIPID RESEARCH CLINICS PROGRAM (1979). The Coronary Primary Prevention Trial: design and implementation. *J. Chronic Dis.* **32**, 609-31.

LIPID RESEARCH CLINICS PROGRAM (1984). The Lipid Research Clinics Coronary Primary Prevention Trial results. I. Reduction in incidence of coronary heart disease. *J. Am. Medical Assoc.* **251**, 351-64.

LUENBERGER, D. G. (1984). *Linear and Nonlinear Programming*, 2nd ed. Reading Mass: Addison-Wesley.

MANTEL, N. & STABLEIN, D. M. (1988). The crossing hazard function problem. *Statistician* **37**, 59-64.

PETO, R. & PETO, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *J. R. Statist. Soc.* A **135**, 185-206.

PHYSICIANS' HEALTH STUDY STEERING COMMITTEE (1983). Physicians' Health Study Protocol. Brookline Mass: Harvard Medical School, Department of Medicine.

SELF, S., PRENTICE, R., IVERSON, D., HENDERSON, M., THOMPSON, D., BYAR, D., INSULL, W., GORBACH, S. L., CLIFFORD, C., GOLDMAN, S., URBAN, N., SHEPPARD, L. & GREENWALD, P. (1988). Statistical design of the Women's Health Trial. *Controlled Clin. Trials* **9**, 119-36.

TARONE, R. E. & WARE, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156-60.

WU, M., FISHER, M. & DEMETS, D. (1980). Sample sizes for long-term medical trials with time-dependent dropout and event rates. *Controlled Clin. Trials* **1**, 109-21.

[*Received January* 1989. *Revised March* 1990]