Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

R. D. Gill & M. Schumacher

A Simple Test of the Proportional Hazards Assumption

Department of Mathematical Statistics Report MS-R8504 July

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

1980 Mathematics subject classification: 62P10, 62905

A simple test of the proportional Hazards Assumption

by

R. Gill
Centrum for Mathematics and Computer Science
Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

&

M. Schumacher
Department of Statistics, University of Dortmund
Postfach 50 05 00, D-4600 Dortmund 50, F.R.G.

* This work has been partially supported by the Deutsche Forschungsgemeinschaft.

1. INTRODUCTION

When comparing two samples of possibly censored survival times it is very often important to assess the proportionality of the underlying hazard functions. This is also true in the more general framework of Cox's proportional hazards model (Cox, 1972) relating the survival time of an individual to other characteristics. Although this assumption of proportionality might be regarded to be only of technical relevance in such a model it is indeed in many applications - at least in the medical field - of substantial importance. So in a study of cancer patients it is of great interest whether the prognostic relevance of the stage of disease at the time of diagnosis can be established over the whole time period. An example where the prognostic relevance is "washed out" in the long term is given by Pocock, Gore & Kerr (1982).

Similarly, in a controlled clinical trial it is very important to distinguish between the kind of uniform superiority of one treatment over another which can be described by a constant relative risk or hazard ratio and a superiority of a treatment which is only of short-term nature.

In order to check the assumption of proportional hazards graphical methods and several test procedures have been proposed so far. For a review of existing methods we refer to Kay (1984) and - restricted to the test procedures - to the examples in chapter 5 of this paper. Nearly all of these tests, however, are based on an arbitrarily chosen partition of the time axis and/or are difficult to compute.

The key idea behind the test procedures proposed in this paper is the observation that in nonproportional hazards situations different two-sample tests, e.g. the logrank and a generalized Wilcoxon test, might come up with very different answers. Our test procedures use this discrepancy as a check of the proportional hazards assumption and are based on the relationship between generalized linear rank tests and estimates of the proportionality constant (Andersen, 1983). This implies that the test statistics can be interpreted in a very natural

way and almost all computational effort has to be done anyway. In addition, a related graphical method is presented which was originally proposed by Lee & Pirie (1981) for comparing trends in series of events.

2. STATISTICAL MODEL AND CONSTRUCTION OF THE TEST STATISTIC

We consider a two-sample censored data situation with sample sizes n_1 and n_2, $n = n_1 + n_2$. More precisely:
Let X_{jk} $(j = 1,2;\ k = 1,\ldots,n_j)$ be independent positive random variables representing survival times which might be times to death, failure or some other well-defined event.
We assume that the distribution functions F_j of X_{jk} are absolutely continuous. Furthermore, the X_{jk} are censored on the right by independent positive random variables C_{jk} which are also independent of the X_{jk}. (Actually much more general censoring models are covered by the mathematical techniques given in the appendix.) Thus, we can only observe

$$\tilde{X}_{jk} = \min(X_{jk}, C_{jk})$$

and

$$\Delta_{jk} = \begin{cases} 1 , & X_{jk} \leq C_{jk} \\ 0 , & X_{jk} > C_{jk} \end{cases} .$$

Usually, one is interested in the testing problem

$$\text{vs.} \begin{cases} H_0^{=} : F_1 = F_2 \\ H_1^{=} : F_1 \neq F_2 \end{cases}$$

whether or not the distributions of the survival times are equal in both groups. The standard statistical methods are the so-called "generalized linear rank tests". Let

$$N_j(t) = \#\text{ deaths (or failures) in group } j \text{ before or at } t \quad (j = 1,2)$$
$$= \#\{k : \tilde{X}_{jk} \leq t,\ \Delta_{jk} = 1\}$$

$$Y_j(t) = \#\text{ at risk in group } j \text{ at } t-$$
$$= \#\{k : \tilde{X}_{jk} \geq t\}$$

then the test statistics can be written as follows

$$Q_K = \int_0^\tau K(t) \left(\frac{dN_2(t)}{Y_2(t)} - \frac{dN_1(t)}{Y_1(t)} \right)$$

where $K(t)$ is a predictable random weight function, i.e. $K(t)$ depends

only on the observations up to t-, and τ is the upper limit of observable survival times. The quantities $\frac{dN_j(t)}{Y_j(t)}$ can be considered as estimators of the hazard functions

$$\lambda_j(t) = \frac{\frac{d}{dt} F_j(t)}{1-F_j(t)}$$

in the two groups, and therefore

$$\hat{\Lambda}_j(t) = \int_0^t \frac{dN_j(s)}{Y_j(s)}$$

as estimators of the cumulative hazard functions

$$\Lambda_j(t) = -\log(1 - F_j(t)) \quad ; \quad j = 1,2$$

(Nelson, 1969).

One is often interested in the special alternative of proportional hazards, i.e., the ratio $\lambda_2(t)/\lambda_1(t)$ is equal to some positive quantity θ. This quantity θ is usually referred to as the "relative risk" and can be simply estimated by the "generalized rank estimator"

$$\hat{\theta}_K := \int_0^\tau K(t)\, d\hat{\Lambda}_2(t) \,/\, \int_0^\tau K(t)\, d\hat{\Lambda}_1(t)$$

(Begun & Reid, 1983; Andersen, 1983). It should be noted that there is a relationship between $\hat{\theta}_K$ and a generalized linear rank test statistic Q_K both having the same weight function K(t), namely

$$(\hat{\theta}_K - 1) \,/\, \sqrt{\widehat{\mathrm{var}}(\hat{\theta}_K)} = Q_K \,/\, \sqrt{\widehat{\mathrm{var}}(Q_K)}$$

(Andersen, 1983). Though many of the generalized linear rank tests were not constructed with the proportional hazards model in mind, they may all be considered as tests based on an estimate for the proportionality constant. In this paper, we are interested in testing the actual proportionality of the hazard functions, i.e. the test problem is given by

$$\begin{cases} H_0 : \lambda_2(t)/\lambda_1(t) = \theta & \text{for some positive } \theta \\ \text{vs.} \\ H_1 : \lambda_2(t)/\lambda_1(t) \neq \theta & \text{for any positiv } \theta \quad . \end{cases}$$

Under H_0 the estimator $\hat{\theta}_K$ converges in probability to θ as sample sizes converge to infinity provided that some technical conditions are satisfied. This implies that, at least for large sample sizes, the difference

between $\hat{\theta}_{K_1}$ and $\hat{\theta}_{K_2}$ for two different weight functions $K_1(t)$ and $K_2(t)$ should be close to zero since both converge to the same quantity θ. For example, we might choose $K_1(t)$ as Gehan's weight function

$$K^{(\text{Gehan})}(t) = Y_1(t) \cdot Y_2(t)$$

(Gehan, 1965) and $K_2(t)$ as the weight function corresponding to the logrank test, namely

$$K^{(\text{Logrank})}(t) = Y_1(t) \cdot Y_2(t) / (Y_1(t) + Y_2(t)) \quad .$$

In this case $K_1(t)$ gives more weight to the early deaths than $K_2(t)$ and $K_2(t)$ weights the late deaths more heavily than $K_1(t)$. Under H_0, however, the estimators for the relative risk based on $K_1(t)$ and $K_2(t)$ should be nearly the same.

On the other hand, these two estimators should be substantially different under H_1, i.e. when the hazard ratio $\lambda_2(t)/\lambda_1(t)$ varies with time and, especially, when the hazard ratio is monotone increasing or decreasing. Thus, in principle, we will base a test statistic on the difference between two generalized rank estimators, $\hat{\theta}_{K_1}$ and $\hat{\theta}_{K_2}$, with two different weight functions.

So consider two weight functions $K_1(t)$ and $K_2(t)$ which are predictable processes and which satisfy $K_j(t) = 0$ when the number in either sample at risk at time t-, $Y_i(t)$ $(i = 1,2)$, is equal to zero. As we have seen above the estimators for the relative risk can be written as

$$(1) \qquad \hat{\theta}_{K_i} = \hat{K}_{i2}/\hat{K}_{i1} \qquad (i = 1,2)$$

where

$$(2) \qquad \hat{K}_{ij} = \int_0^\tau K_i(t)\, d\hat{\Lambda}_j(t) \qquad (j = 1,2) \quad .$$

Instead of using the difference

$$\hat{\theta}_{K_2} - \hat{\theta}_{K_1} = \hat{K}_{22}/\hat{K}_{21} - \hat{K}_{12}/\hat{K}_{11}$$

we can consider the symmetrized version obtained by multiplying by $\hat{K}_{11}\hat{K}_{21}$

$$(3) \qquad Q_{K_1K_2} = \hat{K}_{11}\hat{K}_{22} - \hat{K}_{21}\hat{K}_{12}$$

which should be also close to zero under H_0 .

With arguments which are outlined in detail in the appendix and under certain conditions one can show that under the null hypothesis the asymptotic variance of $Q_{K_1K_2}$ can be consistently estimated by

$$(4) \qquad \widehat{var}(Q_{K_1K_2}) = \hat{K}_{21}\hat{K}_{22}\hat{V}_{11} - \hat{K}_{21}\hat{K}_{12}\hat{V}_{12} - \hat{K}_{11}\hat{K}_{22}\hat{V}_{21} + \hat{K}_{11}\hat{K}_{12}\hat{V}_{22}$$

where

$$(5) \qquad \hat{V}_{ii'} = \int_0^\tau K_i(t)K_{i'}(t) \frac{d(N_1(t)+N_2(t))}{Y_1(t)\cdot Y_2(t)}$$

$$= \int_0^\tau K_i(t)K_{i'}(t) \left[\frac{d\hat{\Lambda}_1(t)}{Y_2(t)} + \frac{d\hat{\Lambda}_2(t)}{Y_1(t)}\right] .$$

In addition, (again under the null hypothesis), the standardized test statistic

$$(6) \qquad T_{K_1K_2} = Q_{K_1K_2} \Big/ (\widehat{var}\ Q_{K_1K_2})^{1/2}$$

has asymptotically ($n\to\infty$) a standard normal distribution. (Note that this statistic is antisymmetric under exchange of K_1 and K_2 or of sample 1 and sample 2. In some situations, another variance estimator may be preferred - see appendix and section 6 - which does not have this property.)

Thus, a two-sided level-α-test can be performed by comparing the absolute value of the test statistic with the $\alpha/2$-fractile of the standard normal distribution. It will be shown later on that this test is consistent against alternatives with a monotone increasing or decreasing hazard ratio, provided the ratio of the two weight functions, $K_2(t)/K_1(t)$, is monotone, too.

Note that $\hat{V}_{ii'}$ is the usual variance estimator of a two-sample test with weight function $\sqrt{K_i(t)K_{i'}(t)}$. This is why we prefer the particular choice of variance estimator given in (4). A disadvantage of this choice is that the estimate can be negative especially when we are far from the null hypothesis. We have already mentioned the possible choice of weight functions $K_1(t) = K^{(Gehan)}(t)$ and $K_2(t) = K^{(Logrank)}(t)$. Another choice for the weight function $K_1(t)$ could be from the class of weight functions $K^{(\rho)}(t)$ proposed by Fleming and Harrington (1982) where

$$K^{(\rho)}(t) = \frac{Y_1(t) \cdot Y_2(t)}{Y_1(t) + Y_2(t)} \; [\hat{S}(t)]^{\rho} \quad .$$

$\hat{S}(t)$ denotes the Kaplan-Meier estimator (Kaplan & Meier, 1958) of the survival function in the combined sample. Obviously, the weight function corresponding to the logrank test is contained in this class ($\rho=0$); the choice $\rho=1$ yields the Peto-Prentice version of the generalized Wilcoxon test (Peto & Peto, 1972; Prentice, 1978). This generalization of the Wilcoxon test is preferable in some respects to Gehan's version; in particular, when censoring is very heavy or when there are different censoring patterns in the two groups (Prentice & Marek, 1979). For a general discussion of this problem see Leurgans (1983). We shall later see that also in our context Prentice's version of the generalized Wilcoxon test is to be preferred. However, with Gehan's version the statistic is easier to compute. Note that the ratio of any two members of this class of weight functions is monotone. The ratio of $K^{(Gehan)}$ with $K^{(0)}$ or with $K^{(1)}$ is monotone, too.

For sake of completeness a proposal due to Fleming, O'Fallon, O'Brien and Harrington (1980) should be mentioned, too. These authors propose in the context of generalized Kolmogorov-Smirnov tests a weight function which is substantially equal to

$$K^{(Fleming)}(t) = \left[\frac{Y_1(t) \cdot Y_2(t)}{Y_1(t) + Y_2(t)}\right]^{1/2} [\hat{S}(t)]^{-1/2}$$

where $\hat{S}(t)$ again denotes the Kaplan-Meier estimator of the survivor function in the combined sample.

The different weight functions mentioned here are displayed and illustrated in one of the examples presented in chapter 5 below.

3. A RELATED GRAPHICAL METHOD

In order to describe the difference between two survival distributions, Lee & Pirie (1981) proposed in another context the so-called "trend function"

$$\gamma(u) = \Lambda_2(\Lambda_1^{-1}(u)) \quad , \quad u \in [0, \Lambda_1(\tau)] \quad .$$

This function has some nice properties: it is a straight line through the origin with slope $\Lambda_2(\tau)/\Lambda_1(\tau)$ when the hazard ratio is constant, it is convex (concave) when the hazard ratio is monotone increasing (decreasing). This is because its derivative, γ', is directly connected with the hazard ratio, namely

$$\gamma'(\Lambda_1(t)) = \lambda_2(t)/\lambda_1(t) \ .$$

Thus, a graphical check on the shape of the hazard ratio can be done by plotting $\hat{\Lambda}_2(t)$ vs. $\hat{\Lambda}_1(t)$.
This graphical method can be easily generalized by using weighted cumulative hazard functions

$$\Lambda_j^{(K)}(t) = \int_0^t K(s)\, d\,\Lambda_j(s) \qquad j = 1,2 \ ,$$

where $K(t)$ is some positive weight function. Then, the trend function is defined by

$$\gamma^{(K)}(u) = \Lambda_2^{(K)}(\Lambda_1^{(K)-1}(u)) \ , \quad u \in [0, \Lambda_1^{(K)}(\tau)] \quad .$$

$\gamma^{(K)}(u)$ has the same nice properties as $\gamma(u)$ and can be estimated by the empirical trend function

$$\hat{\gamma}^{(K)}(u) = \hat{\Lambda}_2^{(K)}(\hat{\Lambda}_1^{(K)-1}(u))$$

with

$$\hat{\Lambda}_j^{(K)}(t) = \int_0^t K(s)\, \frac{dN_j(s)}{Y_j(s)} = \int_0^t K(s)\, d\,\hat{\Lambda}_j(s) \ , \quad j = 1,2 \ .$$

Thus, the proportionality of the hazard function can be graphically checked by plotting $\hat{\Lambda}_2^{(K)}(t)$ vs. $\hat{\Lambda}_1^{(K)}(t)$ and comparing $\hat{\gamma}_K(u)$ with a straight line through the origin with slope $\hat{\Lambda}_2^{(K)}(\tau)/\hat{\Lambda}_1^{(K)}(\tau)$. The signed area between these two functions weighted by some weight function seems

to be an appropriate measure for such a comparison (Figure 1), especially if one is interested in discovering convexity of concavity of $\gamma^{(K)}$. Such a measure can be represented as

$$(7) \qquad \int_0^{\hat{\Lambda}_1^{(K)}(\tau)} \left[\hat{\gamma}^{(K)}(u) - \frac{\hat{\Lambda}_2^{(K)}(\tau)}{\hat{\Lambda}_1^{(K)}(\tau)} u \right] d J(\hat{\Lambda}_1^{(K)-1}(u))$$

where dJ(t) denotes such a weight function. This expression can be rewritten as

$$(8) \qquad \int_0^{\tau} \left[\hat{\Lambda}_2^{(K)}(t) - \frac{\hat{\Lambda}_2^{(K)}(\tau)}{\hat{\Lambda}_1^{(K)}(\tau)} \hat{\Lambda}_1^{(K)}(t) \right] dJ(t)$$

$$= \left[\hat{\Lambda}_2^{(K)}(t) - \frac{\hat{\Lambda}_2^{(K)}(\tau)}{\hat{\Lambda}_1^{(K)}(\tau)} \hat{\Lambda}_1^{(K)}(t) \right] J(t) \Big|_0^{\tau} -$$

$$- \int_0^{\tau} J(t) d \left[\hat{\Lambda}_2^{(K)}(t) - \frac{\hat{\Lambda}_2^{(K)}(\tau)}{\hat{\Lambda}_1^{(K)}(\tau)} \hat{\Lambda}_1^{(K)}(t) \right]$$

$$= -\left[\int_0^{\tau} J(t) d\hat{\Lambda}_2^{(K)}(t) - \frac{\hat{\Lambda}_2^{(K)}(\tau)}{\hat{\Lambda}_1^{(K)}(\tau)} \int_0^{\tau} J(t) d\hat{\Lambda}_1^{(K)}(t) \right]$$

$$= -\left[\int_0^{\tau} J(t)K(t) d\hat{\Lambda}_2(t) - \frac{\int_0^{\tau} K(t) d\hat{\Lambda}_2(t)}{\int_0^{\tau} K(t) d\hat{\Lambda}_1(t)} \int_0^{\tau} J(t)K(t) d\hat{\Lambda}_1(t) \right] .$$

Putting $K(t) = K_2(t)$ and $J(t) = K_1(t)/K_2(t)$ where $K_1(t)$ and $K_2(t)$ are the weight functions used in chapter 2 this expression reduces to

$$(9) \qquad -\left[\hat{K}_{12} - \frac{\hat{K}_{22}}{\hat{K}_{21}} \hat{K}_{11} \right] = Q_{K_1K_2}/\hat{K}_{21} ,$$

an equivalent version of our test statistic $Q_{K_1K_2}$.

The empirical trend function, $\hat{\gamma}_K^{(K)}(u)$, might be used in many other ways, too. The techniques of the appendix can be used to derive the limiting distribution of this function.

4. REMARKS ON CHOICE OF APPROPRIATE WEIGHT FUNCTIONS

The aim of this section is to provide advice on how in practice the random weight functions K_1 and K_2 should be chosen. We base this advice on mathematical considerations of consistency and asymptotic relative efficiency, and on pragmatic considerations of computational convenience. Especially the topic of asymptotic relative efficiency is very technical so we only very briefly sketch the important results here.

Define $K = K_2$ and $J = K_1/K_2$ as in section 3. For large sample results we suppose that as n, the combined sample size increases, the two weight functions converge in probability to deterministic functions, while the number at risk in each sample at each time instant, divided by n, also converges in probability:

$$(10) \qquad Y_i^{(n)}(t)/n \xrightarrow{P} y_i(t)$$

$$K_i^{(n)}(t) \xrightarrow{P} k_i(t)$$

for each t. Define $k(t) = k_2(t)$ and $j(t) = k_1(t)/k_2(t)$. In fact we need slightly strengthened forms of these conditions, namely convergence uniform in $t \in [0,\tau]$, in probability. Under the usual random censorship model this holds for the Y_j's and all the usual weight functions by the Glivenko-Cantelli theorem and its analogue for the product-limit estimator. (Actually for some applications, see section 6 and the appendix, it is useful to replace n in the denominator of (10) by $n_1n_2/(n_1+n_2)$ or by $a^{(n)}$ for some other sequence $a^{(n)} \to \infty$ as $n \to \infty$.)

For consistency results we consider the case of fixed alternatives, Λ_1 and Λ_2 fixed and n o t proportional; for efficiency results we consider a sequence of (non-proportional) alternatives $\Lambda_1^{(n)}$ and $\Lambda_2^{(n)}$ which converge to a proportional hazards situation $\Lambda_2 = \theta\Lambda_1$ at the rate $1/\sqrt{n}$. Consider first the fixed alternatives case. We rewrite the standardized test statistic as

$$T_{K_1K_2} = \sqrt{n}\,(Q_{K_1K_2}/\hat{K}_{21}) \cdot (\hat{K}_{21}/\sqrt{n\,\hat{\mathrm{var}}(Q_{K_1K_2})})$$

(cf. section 3, (7), (8) and (9)) and consider the two bracketed terms separately. From the representation of the first term $Q_{K_1K_2}/\hat{K}_{21}$ given

in section 3, we expect that as $n\to\infty$ this quantity will converge in probability to

$$\int_0^{\Lambda_1^{(k)}(\tau)} \left(\gamma^{(k)}(u) - \frac{\Lambda_2^{(k)}(\tau)}{\Lambda_1^{(k)}(\tau)}\, u\right) dj(\Lambda_1^{(k)-1}(u))$$

$$= \int_0^{\sigma} \left(\gamma^{(k)}(u) - \frac{u}{\sigma}\,\gamma^{(k)}(\sigma)\right) dj(\Lambda_1^{(k)-1}(u))$$

where $\sigma = \Lambda_1^{(k)}(\tau)$, $\Lambda_i^{(k)}(t) = \int_0^t k(s)d\Lambda_i(s)$ and $\gamma^{(k)}(u) = \Lambda_2^{(k)}(\Lambda_1^{(k)-1}(u))$. This can be proved under the above mentioned conditions using the techniques of Appendix I. Immediately we draw the following conclusions, assuming throughout that k_1 and k_2 are positve: If λ_2/λ_1 is increasing (decreasing) then $\gamma^{(k)}$ is convex (concave) and hence $\gamma^{(k)}(u) - \frac{u}{\sigma}\gamma^{(k)}(\sigma)$ is negative (positive). If j is increasing (decreasing) then $j(\Lambda_1^{(k)-1})$ generates a positive (negative) measure. Thus if λ_2/λ_1 and j are both monotone increasing or both monotone decreasing the final result is negative. If they are both monotone but in different directions the final result is positive. (The result will actually be strictly positive or strictly negative under weak non-degeneracy conditions which we do not go into here. Obviously the result is zero if the λ_i's or the k_i's are proportional.)

For the other main term $(\hat{K}_{21}/\sqrt{n\ \hat{\text{var}}(Q_{K_1K_2})})$ we note that under the same convergence conditions we can expect (cf. (2), (4) and (5) in section 2)

$$\hat{K}_{ij} \xrightarrow{P} k_{ij} = \int_0^\tau k_i\, d\Lambda_j$$

$$n\,\hat{V}_{ii'} \longrightarrow v_{ii'} = \int_0^\tau k_i k_{i'} \left(\frac{d\Lambda_1}{y_1} + \frac{d\Lambda_2}{y_2}\right)$$

and hence $(\hat{K}_{21}/\sqrt{n\ \hat{\text{var}}\ Q_{K_1K_2}})$ converges in probability to a strictly positive quantity under weak non-degeneracy conditions.

This can again be proved rigorously using the techniques of the appendix. (Actually, under a fixed alternative hypothesis, $n\ \hat{\text{var}}(Q_{K_1K_2})$ may converge to a negative quantity. If this estimate of the variance is negative, one should replace $(\hat{K}_{21}/\sqrt{n\ \hat{\text{var}}\ Q_{K_1K_2}})$ by $+\infty$.)

Combining these two parts gives consistency against alternatives of monotone hazard rates if K_2/K_1 is monotone, too. We note that this is the case when any two of the weight functions common in survival analysis are used: those corresponding to the logrank test, Gehan's generalization of the Wilcoxon test, or Peto & Peto's and Prentice's generalization of the Wilcoxon test. When we note that all the ingredients $\hat{K}_{ij}$ and $\hat{V}_{ii'}$ of our test statistic (except for $\hat{V}_{12} = \hat{V}_{21}$, corresponding to a weight function $\sqrt{K_1 K_2}$) are needed to compute the two-sample tests with weight functions K_1 and K_2, we see that our test statistic is easy to compute and potentially widely useful.

When considering asymptotic relative efficiencies, the situation is rather more complex. Consider again a sequence of models as above indexed by total sample size n, in which (10) still holds but also the hazard rates $\lambda_1^{(n)}(t)$ and $\lambda_2^{(n)}(t)$ vary with n in such a way that as $n \to \infty$, for all t:

$$\lambda_i^{(n)}(t) \to \lambda_i(t) \qquad \text{where} \quad \lambda_2(t) = \theta\lambda_1(t) \ ,$$

$$n^{1/2}\left(\frac{\lambda_2^{(n)}(t)}{\lambda_1^{(n)}(t)} - \theta\right) \to \ell(t) \ .$$

(These conditions will need to be slightly strengthened to produce rigorous proofs. See Gill (1980, § 5.2) for a complete derivation of the analogous results in the ordinary two-sample problem; see also Leurgans (1984)).

Then it turns out that the standardized test-statistic converges in distribution to the $N(\mu,1)$ distribution with

$$\mu = -\theta^{1/2}\frac{\int_0^\tau jk(\ell-\bar{\ell})d\Lambda}{\sqrt{\left\{\int_0^\tau (j-\bar{j})^2 k^2 d\Lambda/y\right\}}}$$

where $\Lambda = \Lambda_1$, $\bar{\ell} = (\int_0^\tau \ell k \, d\Lambda)/(\int_0^\tau k d\Lambda)$, $\bar{j} = (\int_0^\tau jk \, d\Lambda)/(\int_0^\tau k d\Lambda)$ and $y = y_1 y_2/(y_1 + \theta y_2)$. So the asymptotic power when testing one-sided at level α is $1-\Phi(u_\alpha - \mu)$ where Φ is the standard normal distribution function and $\Phi(u_\alpha) = 1-\alpha$.

Before drawing some conclusions from this formula for μ we present a heuristic derivation of it. The numerator is derived from the formula

for $\sqrt{n}(\hat{K}_{11}\hat{K}_{22} - \hat{K}_{21}\hat{K}_{12})$ in which we replace K_1 and K_2 by k_1 and k_2, and $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ by $\Lambda_1^{(n)}$ and $\Lambda_2^{(n)}$ where $d\Lambda_1^{(n)}(s) = d\Lambda(s)$, $d\Lambda_2^{(n)}(s) = (\theta + \frac{\ell(s)}{\sqrt{n}})d\Lambda(s)$. This gives

$$\sqrt{n}\left\{(\int_0^\tau k_1 d\Lambda)(\int_0^\tau k_2\theta(1+\frac{\ell}{\sqrt{n}})d\Lambda) - (\int_0^\tau k_2 d\Lambda)(\int_0^\tau k_1\theta(1+\frac{\ell}{\sqrt{n}})d\Lambda)\right\}$$

$$= \theta\left\{(\int_0^\tau k_1 d\Lambda)(\int_0^\tau \ell k_2 d\Lambda) - (\int_0^\tau k_2 d\Lambda)(\int_0^\tau \ell k_1 d\Lambda)\right\}$$

$$= \theta\left\{(\int_0^\tau jkd\Lambda)(\int_0^\tau \ell k d\Lambda) - (\int_0^\tau kd\Lambda)(\int_0^\tau j\ell k d\Lambda)\right\}$$

$$= -\theta(\int_0^\tau kd\Lambda)(\int_0^\tau jk(\ell-\bar{\ell})d\Lambda) \ .$$

For the denominator we make the same substitutions in $\sqrt{n\ \hat{var}(Q_{K_1K_2})}$, replacing $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ directly by Λ and $\theta\Lambda$ and Y_1/n and Y_2/n by y_1 and y_2. This gives (cf. (4))

$$n\ \hat{var}(Q_{K_1K_2}) \approx (\int_0^\tau k_2 d\Lambda)(\int_0^\tau \theta k_2 d\Lambda)(\int_0^\tau k_1^2 d\Lambda/y)$$

$$- (\int_0^\tau k_2 d\Lambda)(\int_0^\tau k_1\theta d\Lambda)(\int_0^\tau k_1k_2 d\Lambda/y)$$

$$- (\int_0^\tau k_1 d\Lambda)(\int_0^\tau k_2\theta d\Lambda)(\int_0^\tau k_1k_2 d\Lambda/y)$$

$$+ (\int_0^\tau k_1 d\Lambda)(\int_0^\tau k_1\theta d\Lambda)(\int_0^\tau k_2^2 d\Lambda/y)$$

$$= \theta\left\{(\int_0^\tau kd\Lambda)^2(\int_0^\tau j\,k^2 d\Lambda/y) - 2(\int_0^\tau kd\Lambda)(\int_0^\tau jkd\Lambda)(\int_0^\tau jk^2 d\Lambda/y)\right.$$

$$\left. + (\int_0^\tau jkd\Lambda)^2(\int_0^\tau k^2 d\Lambda/y)\right\}$$

$$= \theta(\int_0^\tau dk\Lambda)^2\left\{(\int_0^\tau j^2k^2 d\Lambda/y) - 2\bar{j}(\int_0^\tau jk^2 d\Lambda/y) + \bar{j}^2(\int_0^\tau k^2 d\Lambda/y)\right\}$$

$$= \theta(\int_0^\tau kd\Lambda)^2\left\{\int_0^\tau (j-\bar{j})^2\,k^2 d\Lambda/y\right\} \quad .$$

Combining these expressions leads to the formula for μ.

The formula can clearly be used for rough power calculations in any particular case once hazard rates and censoring distributions can be hypothesized. We note that the "k function" for typical choices of weight functions is given by

$\frac{y_1\ y_2}{y_1+y_2}$	logrank test
$(1-F)\frac{y_1\ y_2}{y_1+y_2}$	Prentice's Wilcoxon generalization

$y_1\ y_2$ Gehan's Wilcoxon generalization

$(1-F)^{\rho}\dfrac{y_1\ y_2}{y_1+y_2}$ Harrington & Fleming's statistics

where in a random censorship model with $n_i/n \to \rho_i \in (0,1)$ as $n\to\infty$, and with censoring distribution G_i in sample i, $y_i = \rho_i(1-G_i)(1-F_i)$ and F is a distribution with hazard rate

$$\frac{y_1}{y_1+y_2}\lambda_1 + \frac{y_2}{y_1+y_2}\lambda_2 \ ;$$

i.e. a time-varying weighted average of the hazard rates in the two samples, where the weights are proportional to the number at risk in each sample at each time.

We can make some recommendations on choice of k and j (i.e. of k_1 and k_2) by seeking to maximize μ^2 for given y_1, y_2, ℓ, λ and θ by choice of k_1 and k_2. We note that

$$\int_0^\tau jk(\ell-\bar{\ell})d\Lambda = \int_0^\tau k(j-\bar{j})(\ell-\bar{\ell})d\Lambda = \int_0^\tau k(j-\bar{j})\ell d\Lambda \ .$$

So defining $k^* = k(j-\bar{j})$ we find

$$\mu^2 = \theta\left(\int_0^\tau k^*\ell d\Lambda\right)^2 / \left(\int_0^\tau k^{*2}d\Lambda/y\right) \ .$$

Now k* must satisfy $\int_0^\tau k^* d\Lambda = 0$. We therefore consider the problem: maximize $\left(\int_0^\tau k^*\ell d\Lambda\right)$ subject to the constraints $\int_0^\tau k^*\, d\Lambda/y = \text{constant}$, $\int_0^\tau k^* d\Lambda = 0$. By standard methods (cf. Gill, 1980, Lemma 5.2.1) we obtain that $k^* \propto y(\ell-\bar{\ell})$ is the solution. Since $k^* = (j-\bar{j})k$ this suggests that j and k should be chosen with $j \propto \ell$, $k \propto y$. Thus the ratio of the two given weight functions defines the alternatives at which power is maximal (for test-statistics in our class), provided that one of the weight functions has $k \propto y = y_1y_2/(y_1 + \theta y_2)$.

We see that at $\theta=1$, the "optimal k" coincides with the k of the logrank test. We also see that, restricting the weight functions to the above list, taking one of them to be the Gehan weight function means that j will depend strongly on the censoring distributions which seems not a desirable property.

Taking the logrank test and Prentice's Wilcoxon generalization gives a statistic with a nice optimum property at $\theta=1$ and $y_1 \propto y_2$; i.e. we suppose we are close to *equal* hazard rates in the two samples and equal censoring distributions. We then have a statistic which is best, in our class, for testing against alternatives with $\ell \propto (1-F)$. We note that for the logistic location family $F(x;\phi) = (1+e^{-x-\phi})^{-1}$ we have $\frac{\partial}{\partial\phi} \log \lambda(x;\phi) = -(1-F(x;\phi))$. Thus this choice is optimal within our class of statistics, at such a point, for testing proportional hazards versus logistic location alternatives (or any monotone transformation of the latter, since the statistic is a rank statistic). This property is of course related to the fact that the logrank test and Prentice's Wilcoxon test have certain optimality properties for the two-sample proportional hazards and logistic location families, respectively.

In general, however, a test statistic of our class designed with a special alternative in mind is going to be complicated in practice to use. For instance the fact that we would like to have $k \propto y$ suggests taking $K = Y_1Y_2/(Y_1 + \hat{\theta}Y_2)$ where $\hat{\theta}$ is some preliminary estimate of relative risk, e.g. $\hat{\theta}_{\text{logrank}}$, see (1). This K is not predictable but it can be verified that, with some more effort, the same results hold for it as for $K = Y_1Y_2/(Y_1 + \theta Y_2)$. Note that for these K's, $\hat{\theta}_K$ is an efficient estimator of θ under the null-hypothesis of proportional hazards.

Our recommendation therefore is to use the logrank and Prentice's Wilcoxon weight functions as giving a test which is very easy to use and does have some nice optimality properties. Only use the combination of logrank and Gehan's Wilcoxon if atmost simplicity is the aim. Never use the combination of Gehan's and Prentice's Wilcoxon for which the j function only depends on the censoring distributions and is actually constant when there is no censoring, giving a test with no power at all.

The above efficiency calculations were made *within* our class of statistics. It seems likely that the best statistic in our class for a particular alternative will have some global optimality property among all tests in a wider class (e.g. all rank tests), but we have not investigated this in detail yet. The analogue question for ordinary censored data linear rank tests still needs thorough investigation.

5. EXAMPLES

Fleming, O'Fallon, O'Brien and Harrington (1980) present data on time from treatment to progression of disease of 35 patients with stage II or IIA ovarian cancer. The data of these patients who were treated in the Mayo Clinic are listed in table 1.

	times from treatment to disease progression
Stage II patients ($n_1 = 15$)	28, 89, 175, 195, 309, 377^+, 393^+, 421^+, 447^+, 462, 709^+, 744^+, 770^+, 1106^+, 1206^+
Stage IIA patients ($n_2 = 20$)	34, 88, 137, 199, 280, 291, 299^+, 300^+, 309, 351, 358, 369, 369, 370, 375, 382, 392, 429^+, 451, 1119^+

Table 1: Times from treatment to disease progression of patients with ovarian cancer; censored observations are marked with "+"

Figure 2 displays the Nelson estimates of the cumulative hazard functions which show that grade influences the rate of progression only towards the end of the time scale. How the differences between the two hazard functions are weighted by various two-sample tests is shown in figure 3. In order to make the order of magnitude of the different weight functions comparable we have normalized them by dividing by the square root of the corresponding variance estimators.

Giving different weight to the "early" and "late" differences of the two hazard functions results in different P-values for the various two-sample tests considered. In particular, these P-values are 0.018 for the logrank test, 0.047 for Harrington & Fleming's test with $\rho=0.5$, and 0.109 and 0.134 for Prentice's and for Gehan's generalized Wilcoxon test, respectively. The (in many respects) extreme proposal of Fleming et al. (1980) yields a P-value of 0.015. Since Harrington & Fleming's

weight function with $\rho=0.5$ is only a compromise between the logrank test and the Wilcoxon generalizations it seems to be adequate for our purposes to compare the generalized rank estimates for the relative risk based on the logrank weight function and on Gehan's or Prentice's weight functions. For the logrank weight function we obtain $\hat{\theta}^{(\text{Logrank})} = 2.78$ whereas for Gehan's and Prentice's weight function we get $\hat{\theta}^{(\text{Gehan})} = 1.99$ and $\hat{\theta}^{(\text{Prentice})} = 2.02$, respectively. The difference between these values indicates a lack of proportionality. This is established by calculating the test statistics proposed in chapter 2 yielding $T = 2.83$ $(p = 0.005)$ for the Gehan vs. logrank comparison and $T = 2.46$ $(p = 0.014)$ for the Prentice vs. logrank comparison.

All p-values mentioned here are two-tailed. The reason for the similar behaviour of Gehan's and Prentice's weight function is the relatively light censoring in this example. Up to the largest uncensored time, there are only four censored observations among the stage II patients and three censored observations among the stage IIA patients.

Thus, in these ovarian cancer data the null-hypothesis of proportional hazards has to be rejected. This is also visually supported by the plot of the empirical trend function using the logrank weight function as displayed in fig. 4. The data have also been used by Breslow (1984) and Breslow et al. (1984) who calculated a test statistic for "acceleration" based on Cox's (1972) original proposal yielding a p-value of 0.017, in concordance with our results.

The second example is a controlled clinical trial in chronic stable angina comparing the survival times of patients receiving coronary artery bypass graft surgery and of patients receiving a conservative medical treatment.

Details of the trial which was undertaken by the Veterans administration can be found in Detre et al. (1977). A first impression of the results of this trial - the sample sizes are considerable: $n_1 = 507$ and $n_2 = 508$ - may be gained from the display of the hazard ratio in

fig. 5. The hazard ratio has been estimated by assuming a piecewise exponential model. Figure 5 shows that the risk is more than twice as high for the surgically treated patients immediately after treatment. Then the hazard ratio rapidly decreases and finally remains constant at a level of about 0.75 after three years. This should be an excellent example for a nonproportional hazards situation but our test statistic based on a logrank vs. Gehan comparison yields only a value of $T = -1.27$ associated with a nonsignificant p-value of 0.2. Figure 6 displays the values of our standardized test statistic T calculated after 1,2,3,...,8 years. This strongly suggests that departure from the proportionality of the hazard functions is restricted to the first four years after treatment. This is also confirmed by the plot of the empirical trend function based on the logrank weight function which is displayed in figure 7 and which also indicates that the hazard ratio is not monotone in this example.

These data show very clearly the limitations of our proposed test statistic. It is designed only to detect departures from the proportionality of the hazard functions when the hazard ratio is monotone. This has to be seen in contrast to the other "omnibus" test procedures based onan arbitrarily chosen partition of the time axis. The resulting p-values of some of these test procedures (Andersen (1982), Schoenfeld (1980) and Schumacher & Vaeth (1984)) when using a partition of the time axis into nine intervals are given in table 2, all of them leading to a rejection of the null-hypothesis of proportional hazards.

test statistic	p-value
Andersen (Wald)	0.031
Andersen (likelihood ratio)	0.023
Schoenfeld	0.026
Schumacher & Vaeth 1	0.002
Schumacher & Vaeth 2	0.041

Table 2: Results of various test statistics for testing the proportionality of the hazard functions in the Veterans Administration data

A more thorough discussion of these data can be found in Schumacher (1982, 1984).

Two other examples will be mentioned very briefly. The first of these uses data on time to remission of leucemia patients. It was presented by Freireich et al. (1963) and used as an example by very many authors. (e.g. Gehan (1965), Cox (1972), Schoenfeld (1980), Begun & Reid (1984), Nagelkerke et al. (1984), Schumacher & Vaeth (1984), Wei (1984)). As is shown by Begun & Reid (1984) the various estimates of relative risk are not too different compared with their standard errors. Thus it is not astonishing that our test statistic based on a logrank vs. Prentice comparison yields a p-value of 0.72. In this example Gehan's weight function is rather sensitive to the highly unbalanced censoring patterns - there are no censored observations at all in the second sample. The p-value obtained by our test statistic agrees with the p-values obtained by test statistics proposed by the authors mentioned above and are listed in table 3.

test statistic	p-value
Andersen (Wald)	0.26
Andersen (likelihood ratio)	0.26
Nagelkerke, Oostring & Hart	0.65
Schoenfeld	0.41
Schumacher & Vaeth 1	0.53
Schumacher & Vaeth 2	0.55
Wei	0.65

Table 3: Results of various test statistics for testing the proportionality of the hazard functions in the Freireich data.

The last example is based on the canine transplant data presented by Prentice & Marek (1979). This is in many respects a very extreme example because the sample sizes and the censoring patterns in both groups are very different. This is reflected by the large difference between Gehan's and Prentice's weight function yielding a value of $T = 0.09$

for the logrank vs. Prentice and of $T = 3.77$ for the logrank vs. Gehan comparison. A thorough discussion of this phenomenon is given by Prentice & Marek (1979).

6. OTHER APPLICATIONS

6.1 TESTING FOR TREND IN POISSON PROCESSES

Consider two non-homogeneous Poisson processes N_1 and N_2 with intensity functions $\mu_1(t)$ and $\mu_2(t)$, $t \in [0,\tau]$. Lee & Pirie (1981) consider the problem of testing the null-hypothesis $\mu_2(t)/\mu_1(t) = \theta$ versus $\mu_2(t)/\mu_1(t)$ monotone and proposed a graphical technique and a test statistic which, in a certain sense, are special cases of our methods. In fact our initial aim was precisely to investigate whether their methods could be used in the analoguous censored data problem.

If we choose any constant a and define "numbers at risk" processes $Y_i(t)$ and "hazard rates" $\lambda_i(t)$ by $Y_i(t) \equiv a$, $\lambda_i(t) = \mu_i(t)/a$ $(i = 1,2)$ then the processes and functions N_i, Y_i and λ_i share many properties of the same quantities in the censored data problem. In particular all the mathematical results of the appendix apply without any change at all to this new situation.
If we take a=1, $K(t) = 1$ for all t and $J(t) = N_1(t-) + N_2(t-) + 1$ - recall $K(t) = K_2(t)$ and $J(t) = K_1(t)/K_2(t)$ - then the standardized statistic of section 2 becomes asymptotically equivalent (under the null-hypothesis or under a sequence of contiguous alternatives) to the standardized version of Lee & Pirie's (1981) statistic while the plot of section 3 becomes precisely their relative trend plot. Using the alternative variance estimator (A2) with $\alpha_1 = \alpha_2 = 1$ (see appendix) the standardized statistic is actually equal to $\sqrt{R/(R-1)}$ times their statistic, where $R = N_1(\tau) + N_2(\tau)$.
An interesting *difference* between the two statistics is that theirs is proposed as a conditional test, conditional on the values of $N_1(\tau)$ and $N_2(\tau)$, so their standardized test uses a *conditional* variance. Their large sample theory is also theory on asymptotic conditional distributions.

One can investigate large sample properties in exactly the same way as in section 4 (in fact we used the term "asymptotically equivalent" just now in the sense indicated in section 4). We consider a sequence of problems indexed by n in which one observes two Poisson processes over

the same fixed time interval $[0,\tau]$ for all n, but with larger and larger intensity functions $\mu_1^{(n)}$ and $\mu_2^{(n)}$. Now it is useful that we earlier introduced the constant a (which till now we took equal to 1): we let this number depend on n, and suppose that $\mu_1^{(n)}$ and $\mu_2^{(n)}$ grow in such a way that $\lambda_i^{(n)}(t) = \mu_i^{(n)}(t)/a^{(n)} \to \lambda_i(t)$ as $n \to \infty$, $a^{(n)} \to \infty$ as $n \to \infty$. Choosing J and K as above (or rather $J^{(n)}(t) = (N_1^{(n)}(t-) + N_2^{(n)}(t-) + 1)/a^{(n)}$, $K^{(n)}(t) \equiv 1$), we obtain $j(t) = \Lambda_1(t) + \Lambda_2(t)$, $k(t) = 1$ and (replacing n in the denominator of (10) by $a^{(n)}$) $y_i(t) = 1$. So $y = y_1y_2/(y_1 + \theta y_2) = 1(1+\theta) = \text{constant}$. Thus this choice of k and j has some optimality properties when $\ell(t) \propto \Lambda(t)$. This corresponds to parametric alternatives to the proportional intensities model of the form

$$\lambda_2(t;\phi) = \theta\lambda_1(t)\exp(\phi\Lambda_1(t) + \sigma(\phi))$$

as $\phi \to 0$.

6.2 TESTING FOR EXPONENTIALITY VERSUS A MONOTONE HAZARD RATE IN THE ONE-SAMPLE CASE: THE TOTAL TIME ON TEST STATISTIC

A one-sample analogue of our problem is also of very great interest. Suppose we are given a specified hazard rate $\lambda_1(t)$ (e.g. $\lambda_1(t) = 1$ for all t), and a censored sample from a distribution with hazard rate $\lambda_2(t)$. Suppose we wish again to test the hypothesis $\lambda_2(t)/\lambda_1(t) = \theta$ for some constant θ versus the alternative $\lambda_2(t)/\lambda_1(t)$ monotone. In the special case $\lambda_1(t) \equiv 1$ this is the same as testing exponentiality versus alternatives of a monotone hazard rate. The total time on test statistic (see Barlow, Bartholomew, Bremner & Brunk (1972), section 6.2 and Aalen & Hoem (1978), section 3.4) is a well-known statistic for this purpose. (Note that as in section 6.1 we can also consider the analogous problem, for which the total time on test statistic is available, too, on the intensity function of a Poisson process.)

In fact the one-sample analogue of our class of statistics contains the total time on test statistic as a special case. Also the total time on test plot is (up to a scale transformation of each axis) our relative

trend plot with suitable choice of K function.

Moreover our theoretical results (at least, their easier one-sample analogues) provide immediately large sample results for the total time on test statistic with censored data. (Aalen & Hoem (1978) and Barlow & Proschan (1969) both claim to give general results, but in fact in both papers it is tacitly assumed that one stops observation at a predetermined uncensored observation so that the number of uncensored observations is non-random. The statistic was only ever introduced in this situation anyway. Here we suppose that observation stops at a *fixed time*.)

We define our class of one-sample statistics exactly as in section 1, using the index 2 to indicate the sample actually available, and using the index 1 to indicate a fictitious sample from a distribution with the given hazard rate λ_1 which is so large that $\hat{\Lambda}_1$ and Λ_1 are taken to be identical (cf. (2), (3), (4), (5), (6); in the last expression of (5) we take $Y_1(t) \equiv \infty$). For asymptotic results (cf. section 4) we replace n in the denominator of (10) by n_2 and take $y_1(t) = \infty$.

Taking $K(t) = Y_2(t)$, $J(t) = N_2(t-) + 1$ and $\Lambda_1(t) = t$ we obtain a standardized statistic asymptotically equivalent to the standardized total time on test statistic. To show this, let us work rather with the alternative variance estimator (A2) with $\alpha_1 = 0$, $\alpha_2 = 1$ (see appendix). Let $R = N_2(\tau)$ and let $0 < T_1 < T_2 < \ldots < T_R < \tau$ be the ordered uncensored observations in $[0,\tau]$ (i.e. the jump times of $N_2(t)$). Let $T_0 = 0$, $T_{R+1} = \tau$, and define $D_j = \int_{T_{j-1}}^{T_j} Y_2(t)dt$, $j = 1,\ldots,R+1$. First we give an expression for the standardized total time on test statistic (Barlow, Bartholomew, Bremner & Brunk (1972), p. 268) using the data on the time internal $[0,T_r]$; i.e. as if on knew beforehand that there would be at least R=r (not random) uncensored observations, and stopped registering failures and censorings at the time of the r-th failure. The statistic can then be written as

$$\frac{\frac{1}{R-1}\sum_{i=1}^{R-1}\left(\sum_{j=1}^{i} D_j\right) \Big/ \sum_{j=1}^{R} D_j - \frac{1}{2}}{\sqrt{1/(12(R-1))}}$$

$$= \frac{\sum_{j=1}^{R}(R-j)D_j - \frac{1}{2}(R-1)\sum_{j=1}^{R} D_j}{\sum_{j=1}^{R} D_j \sqrt{\frac{1}{12}(R-1)}} = -\frac{\sum_{j=1}^{R} j\, D_j - \frac{1}{2}(R+1)\sum_{j=1}^{R} D_j}{\sum_{j=1}^{R} D_j \sqrt{\frac{1}{12}(R-1)}} \quad .$$

On the other hand, we obtain from (2)

$$\hat{K}_{11} = \int_0^\tau Y_2(t)(N_2(t-)+1)dt = \sum_{j=1}^{R+1} j\, D_j$$

$$\hat{K}_{22} = \int_0^\tau Y_2(t)\frac{dN_2(t)}{Y_2(t)} = R$$

$$\hat{K}_{21} = \int_0^\tau Y_2(t)dt = \sum_{j=1}^{R+1} D_j$$

$$\hat{K}_{12} = \int_0^\tau Y_2(t)(N(t-)+1)\frac{dN_2(t)}{Y_2(t)} = \frac{1}{2}R(R+1)\ ,$$

so that by (3)

$$Q_{K_1K_2} = R\left\{\sum_{j=1}^{R+1} j\, D_j - \frac{1}{2}(R+1)\sum_{j=1}^{R+1} D_j\right\}\ .$$

Also in (A2) with $\alpha_1 = 0$, $\alpha_2 = 1$ we find $\hat{\Lambda}_0 = \hat{\Lambda}_2$, hence $\hat{c}_2 = 1$ and $\hat{c}_1 = \hat{K}_{21}/\hat{K}_{22} = (\sum_{j=1}^{R+1} D_j)/R$. Putting $Y_1(t) = \infty$ this gives

$$\tilde{var}(Q_{K_1K_2}) = \int_0^\tau (\hat{K}_{22}K_1(t) - \hat{K}_{12}K_2(t))^2\,\hat{c}_1^2\frac{dN_2(t)}{Y_2(t)^2}$$

$$= \int_0^\tau (R(N_2(t-)+1) - \frac{1}{2}R(R+1))^2\, dN_2(t)\left(\left(\sum_{j=1}^{R+1} D_j\right)/R\right)^2$$

$$= \int_0^\tau ((N_2(t-)+1) - \frac{1}{2}(R+1))^2\, dN_2(t)\left(\sum_{j=1}^{R+1} D_j\right)^2$$

$$= \left(\sum_{j=1}^{R+1} D_j\right)^2 \frac{1}{12}(R+1)\,R(R+1)\ .$$

Thus the standardized statistic becomes

$$\frac{\sum_{j=1}^{R+1} j\, D_j - \frac{1}{2}(R+1)\sum_{j=1}^{R+1} D_j}{\sum_{j=1}^{R+1} D_j \sqrt{\frac{1}{12}(R-1)(R+1)/R}}\ .$$

This differs from the total time on test statistic by a factor $-\sqrt{\frac{R}{R+1}}$ and by inclusion of an extra term j=R+1 in each summation. The minus sign was to be expected since the one-sided form of the total time on test statistic (reject for large values) is designed against alternatives in which λ_2/λ_1 is increasing. Since $J = K_1/K_2$ is increasing, too, our statistic should take on large negative values under such an alternative.

(K_1 gives more weight to later times, so that $\hat{\theta}_{K_1}$ tends to be larger than $\hat{\theta}_{K_2}$.) The other differences are negligeable for large samples.

Note that taking $K(t) = Y_2(t)$ and $dN_1(t)/Y_1(t) = dt$ our relative trend plot becomes a plot of $\int_0^t Y_2(s)ds$ versus $N_2(t)$, $t \in [0,\tau]$. The total time on test plot based on R observations is a plot of $\int_0^t Y_2(s)ds/\int_0^{T_R} Y_2(s)ds$ against $N_2(t)/R$.

A more thorough discussion of these topics can be found in a separate paper by on of the authors (Gill (1985)).

7. CONCLUDING REMARKS

In a recent paper Wei (1984) proposed another goodness-of-fit test for proportional hazards which is - at least in a wide sense - related to the methods proposed in this paper. In particular, Wei's test can be shown to be asymptotically equivalent to a Kolmogorov-Smirnov-type version of our test statistics using a special weight function. Details have been worked out by Andersen (1983).

A generalization to the case of p-samples of our test statistics is possible in principle but neither simple nor straightforward. The reason for this is based on the fact that when comparing the hazard function of the j-th sample with the hazard function of the pooled other samples, the hazard ratio is no longer proportional even under the null-hypothesis. Thus building up a test statistic in a 'Kruskal-Wallis'-manner - as described by Andersen, Borgan, Gill & Keiding (1982) - is not feasible. A suitable test statistic, however, could be based on all pairwise comparisons. The asymptotic null-hypothesis covariance matrix of the $\frac{1}{2}p(p-1)$ pairwise test statistics using the *same* weight functions K_1 and K_2 in every comparison can be shown to have rank p-1. Thus these pairwise test statistics can be combined to a global test statistic which has asymptotically under the null-hypothesis a χ^2-distribution with p-1 degrees of freedom. We omit the details.

The strengths and weaknesses of out tests are best illustrated by the examples in chapter 5 featuring various practically important situations. Although a theoretical and/or empirical comparison with all the other proposals still has to be done - as also stated by Kay (1984) - the test procedures proposed in this paper provide an attractive tool for assessing the proportionality of hazard functions.

APPENDIX

Here we indicate that the counting process methods of e.g. Gill (1980) or Andersen, Borgan, Gill & Keiding (1982) can be used to derive large sample results about our test statistic and graphical method. In fact we just consider asymptotic normality of the test statistic under the null-hypothesis. Consistency, efficiency and (for the trend function) weak convergence results can be obtained using the same tools without further difficulties. See also Gill (1984) for an informal introduction to these methods. For the trend function one also needs the methods of Vervaat (1972) for dealing with weak convergence of inverse processes. Also all these results are immediately available for a class of counting process models which includes just as a special case the random censorship model.

Consider a bivariate counting process $(N_1,N_2) = ((N_1(t), N_2(t)) : t \in [0,\tau])$ with intensity process $(Y_1\lambda_1, Y_2\lambda_2)$ such that $\lambda_j(t) = \theta_j\lambda(t)$ for all t. So Y_1 and Y_2 are non-negative predictable processes and λ_1 and λ_2 are fixed, proportional non-negative functions. Let K_1 and K_2 be two predictable processes. As in section 1 define

$$\hat{\Lambda}_j(t) = \int_0^t dN_j(s)/Y_j(s) \qquad j = 1,2$$

$$\hat{K}_{ij} = \int_0^\tau K_i(s)d\hat{\Lambda}_j(s)$$

$$\Lambda_j(t) = \int_0^t \lambda_j(s)ds \quad .$$

(We set $Y_j^{-1} = 0$ where $Y_j = 0$.)
Define also

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

$$\bar{K}_{ij} = \int_0^\tau K_i(s)d\Lambda_j(s)$$

$$\bar{K}_i = \int_0^\tau K_i(s)d\Lambda(s) \ ,$$

where we suppose $\Lambda(\tau) < \infty$. We also suppose the sample paths of $|K_i|$ $(i = 1,2)$ and Y_j^{-1} $(j = 1,2)$ are almost surely bounded and that K_1 and K_2

are both zero where Y_1 or Y_2 are.
Considering $\hat{K}$ as the 2×2-matrix with elements $\hat{K}_{ij}$, we can write our test statistic $Q_{K_1K_2}$ (cf. (3)) as

$$Q_{K_1K_2} = \det(\hat{K}) \quad .$$

Note that $\bar{K}_{ij} = \theta_j \bar{K}_i$ and hence $\det(\bar{K}) = 0$.
We wish to derive a large sample result on $Q_{K_1K_2}$ so we consider a sequence of the models described above indexed by n; so N_1, N_2, Y_1, Y_2, K_1 and K_2 all depend on n but λ_1 and λ_2 remain fixed (and proportional). We suppress this dependence on n from our notation.

We recall that a possible estimator $\widehat{\mathrm{var}}(Q_{K_1K_2})$ of the asymptotic null-hypothesis variance of $Q_{K_1K_2}$ is defined by (4) and (5) which we can re-write as

$$\text{(A1)} \quad \widehat{\mathrm{var}}(Q_{K_1K_2}) = \int_0^\tau (\hat{K}_{21}K_1(t) - \hat{K}_{11}K_2(t))(\hat{K}_{22}K_1(t) - \hat{K}_{12}K_2(t)) \left(\frac{d\hat{\Lambda}_1(t)}{Y_2(t)} + \frac{d\hat{\Lambda}_2(t)}{Y_1(t)} \right) .$$

We define a whole class of further possible estimators by, for given $\alpha_1, \alpha_2 \geq 0$, $\alpha_1 + \alpha_2 > 0$ defining $\hat{\Lambda}_0 = \alpha_1\hat{\Lambda}_1 + \alpha_2\hat{\Lambda}_2$. Define $\hat{K}_{ij} = \int_0^\tau K_i \, d\hat{\Lambda}_j$ also for j=0 and let

$$\hat{c}_j = \hat{K}_{2j}/\hat{K}_{20} \quad , \qquad j = 1,2 \quad .$$

Then we set

$$\text{(A2)} \quad \widetilde{\mathrm{var}}(Q_{K_1K_2}) = \int_0^\tau (\hat{K}_{20}K_1(t) - \hat{K}_{10}K_2(t))^2 \, \hat{c}_1\hat{c}_2 \left(\frac{\hat{c}_1}{Y_2(t)} + \frac{\hat{c}_2}{Y_1(t)} \right) d\hat{\Lambda}_0(t) \quad .$$

Theorem

Suppose there exists a sequence $a^{(n)}$, $a^{(n)} \to \infty$ *as* $n \to \infty$, *and fixed functions* y_1, y_2, k_1 *and* k_2 *such that*

$$\sup_{t\in[0,\tau]} |Y_j(t)/a_n^{(n)} - y_j(t)| \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty, \; j = 1,2$$

$$\sup_{t\in[0,\tau]} |K_i(t) - k_i(t)| \xrightarrow{P} 0 \qquad \text{as} \quad n \to \infty, \; i = 1,2$$

where $|k_i|$ $(i = 1,2)$ *and* y_j^{-1} $(j = 1,2)$ *are bounded on* $[0,\tau]$. *Then as* $n \to \infty$

$$(a^{(n)})^{1/2} \, Q_{K_1K_2} \xrightarrow{D} N(0,\sigma^2)$$

$$a^{(n)}\ \widehat{var}(Q_{K_1K_2}) \xrightarrow{\mathcal{P}} \sigma^2 ,$$

and (for any α_1, α_2*)*

$$a^{(n)}\ \widetilde{var}(Q_{K_1K_2}) \xrightarrow{\mathcal{P}} \sigma^2 ,$$

where

$$\sigma^2 = \int_0^\tau (\bar{k}_2\ k_1(t) - \bar{k}_1\ k_2(t))^2\ \theta_1\theta_2\left(\frac{\theta_1}{y_2(t)} + \frac{\theta_2}{y_1(t)}\right) d\Lambda(t)$$

and $\bar{k}_i = \int_0^\tau k_i(t)d\Lambda_i(t)$.

Before proving the theorem we give, as a lemma, a version of the δ-method, which will enable us to derive asymptotic normality of $a^{(n)\frac{1}{2}}(\det(\hat{K}) - \det(\bar{K}))$ from asymptotic normality of $a^{(n)\frac{1}{2}}(\hat{K} - \bar{K})$.

Lemma

Let $\hat{X}^{(n)}$, $\bar{X}^{(n)}$ *be random column-vectors in* $\mathbb{R}^p$ *and* μ *a fixed vector. Let* $f : \mathbb{R}^p \to \mathbb{R}$ *be differentiable in a neighbourhood of* μ *with derivative* f *which is continuous at* μ. *Suppose for some numbers* $a^{(n)} \to \infty$ *as* $n \to \infty$ *and a random vector* Z *we have as* $n \to \infty$

$$a^{(n)\frac{1}{2}}\ (\hat{X}^{(n)} - \bar{X}^{(n)}) \xrightarrow{\mathcal{D}} Z ,$$

$$\bar{X}^{(n)} \xrightarrow{\mathcal{P}} \mu$$

(and hence also $\hat{X}^{(n)} \xrightarrow{\mathcal{P}} \mu$*). Then*

$$a^{(n)\frac{1}{2}}\ (f(\hat{X}^{(n)}) - f(\bar{X}^{(n)})) \xrightarrow{\mathcal{D}} f(\mu)^T Z \quad \text{as} \quad n \to \infty.$$

Proof:

By the mean value theorem we have (with probability converging to 1 as $n \to \infty$) that

$$a^{(n)\frac{1}{2}}\ (f(\hat{X}^{(n)}) - f(\bar{X}^{(n)})) = f(\tilde{X}^{(n)})\ (a^{(n)\frac{1}{2}}(\hat{X}^{(n)} - \bar{X}^{(n)}))$$

where $\tilde{X}^{(n)}$ lies on the line segment between $\hat{X}^{(n)}$ and $\bar{X}^{(n)}$ in $\mathbb{R}^p$. The result is now obvious. ∎

Proof of the theorem:

By a routine application of counting process methods it is easy to verify that for $i,j = 1,2$ we have (jointly)

$$a^{(n)1/2} Z_{ij}^{(n)} = a^{(n)1/2} \int_0^\tau K_i(t)(d\hat{\Lambda}_j(t) - d\Lambda_j(t))$$

$$= a^{(n)1/2} \int_0^\tau K_i(t) \frac{(dN_j(t) - Y_j(t)d\Lambda_j(t))}{Y_j(t)}$$

$$\xrightarrow{\mathcal{D}} \int_0^\tau k_i(t)dW_j(t)$$

where W_1 and W_2 are independent Gaussian processes with zero means, independent increments, and variance functions $\mathrm{var}(W_j(t)) = \int_0^\tau d\Lambda_j(s)/y_j(s)$. So applying the lemma with $\hat{X}^{(n)}$, $\bar{X}^{(n)}$ and f replaced by $\hat{R}$, $\bar{K}$ and $\det(\cdot)$ we obtain

$$a^{(n)1/2} Q_{K_1K_2} \xrightarrow{\mathcal{D}} \sum_{i,j} \bar{k}^{ij} \int_0^\tau k_i(t)dW_j(t)$$

where $\bar{k}^{ij} = (-1)^{i+j} \bar{k}_{3-i,3-j}$; $i,j = 1,2$; $\bar{k}_{ij} = \int_0^\tau k_i(t)d\Lambda_j(t)$. Now $\bar{k}_{ij} = \theta_j \bar{k}_i$ and

$$\sum_{i,j} \bar{k}^{ij} \int_0^\tau k_i(t)dW_j(t) = \int_0^\tau (\bar{k}_{22}k_1(t) - \bar{k}_{12}k_2(t))dW_1(t)$$

$$+ \int_0^\tau (-\bar{k}_{21}k_1(t) + \bar{k}_{11}k_2(t))dW_2(t)$$

$$\overset{\mathcal{D}}{=} N(0,\sigma^2)$$

where

$$\sigma^2 = \int_0^\tau (\bar{k}_{22}k_1(t) - \bar{k}_{12}k_2(t))^2 \, d\Lambda_1(t)/y_1(t)$$

$$+ \int_0^\tau (\bar{k}_{21}k_1(t) - \bar{k}_{11}k_2(t))^2 \, d\Lambda_2(t)/y_2(t)$$

$$= \int_0^\tau \theta_2^2\theta_1(\bar{k}_2k_1(t) - \bar{k}_1k_2(t))d\Lambda(t)/y_1(t)$$

$$+ \int_0^\tau \theta_1^2\theta_2(\bar{k}_2k_1(t) - \bar{k}_1k_2(t))d\Lambda(t)/y_2(t)$$

(A3)
$$= \int_0^\tau (\bar{k}_2k_1(t) - \bar{k}_1k_2(t))^2 \, \theta_1\theta_2\left(\frac{\theta_1}{y_2(t)} + \frac{\theta_2}{y_1(t)}\right) d\Lambda(t)$$

(A4)
$$= \int_0^\tau (\bar{k}_{21}k_1(t) - \bar{k}_{11}k_2(t))(\bar{k}_{22}k_1(t) - \bar{k}_{12}k_2(t))\left(\frac{d\Lambda_1(t)}{y_2(t)} + \frac{d\Lambda_2(t)}{y_1(t)}\right) .$$

This proves the first part of the theorem on asymptotic normality. For the second part on consistency of the variance estimators we note first that again by routine methods $a^{(n)}\hat{\text{var}}(Q_{K_1K_2})$ (cf. (A1)) converges in probability to the expression for σ^2 given by (A4). Next, by multiplying Λ by a constant if necessary (and dividing θ_1 and θ_2 by the same constant) we can identify Λ and $\Lambda_0 = \alpha_1\Lambda_1 + \alpha_2\Lambda_2$ (for any given choice of α_1 and α_2). We now note that $\hat{K}_{ij} \xrightarrow{\mathcal{P}} \bar{k}_{ij}$ for $i = 1,2$; $j = 0,1,2$ where $\bar{k}_{i0} = \bar{k}_i$. So $\hat{c}_j \xrightarrow{\mathcal{P}} \theta_j$ for $j = 1,2$. It is now also easy to see that $\tilde{\text{var}}(Q_{K_1K_2})$ given by (A2) converges in probability to the equivalent expression (A3) for σ^2. Δ

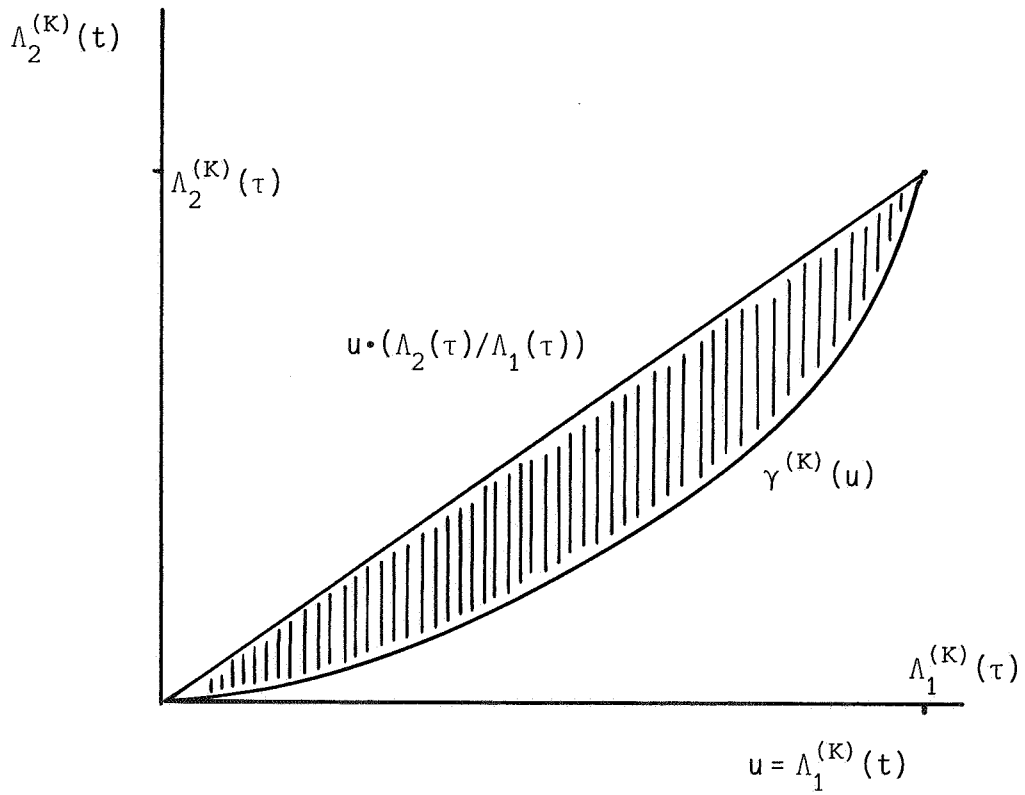

Figure 1: Comparison of $\gamma^{(K)}(u)$ and the straight line $u \cdot (\Lambda_2(\tau)/\Lambda_1(\tau))$ as graphical check for the proportional hazards assumption.

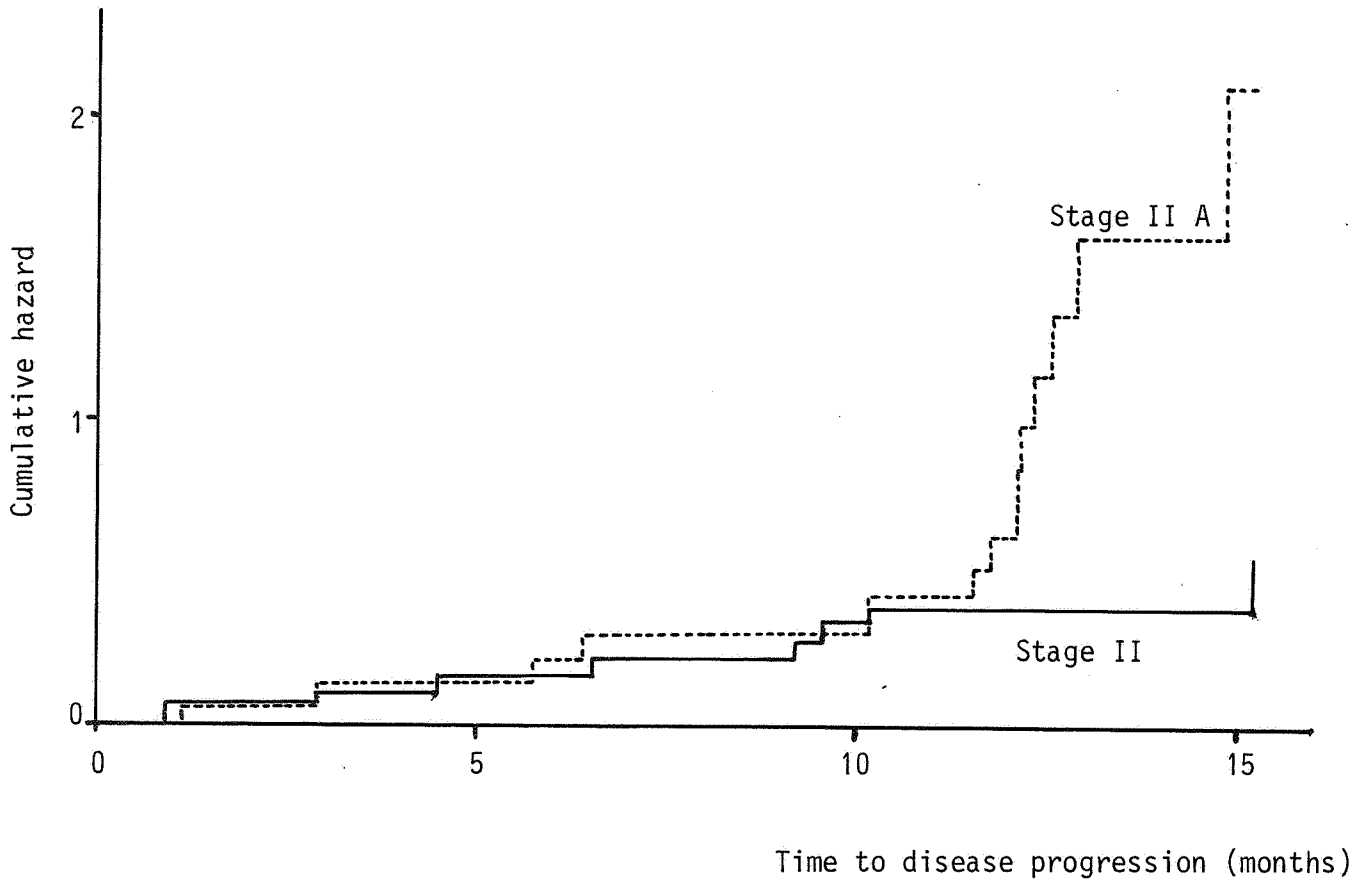

Figure 2: Empirical cumulative hazard functions of patients with stage II (——) and stage II A (---) ovarian cancer.

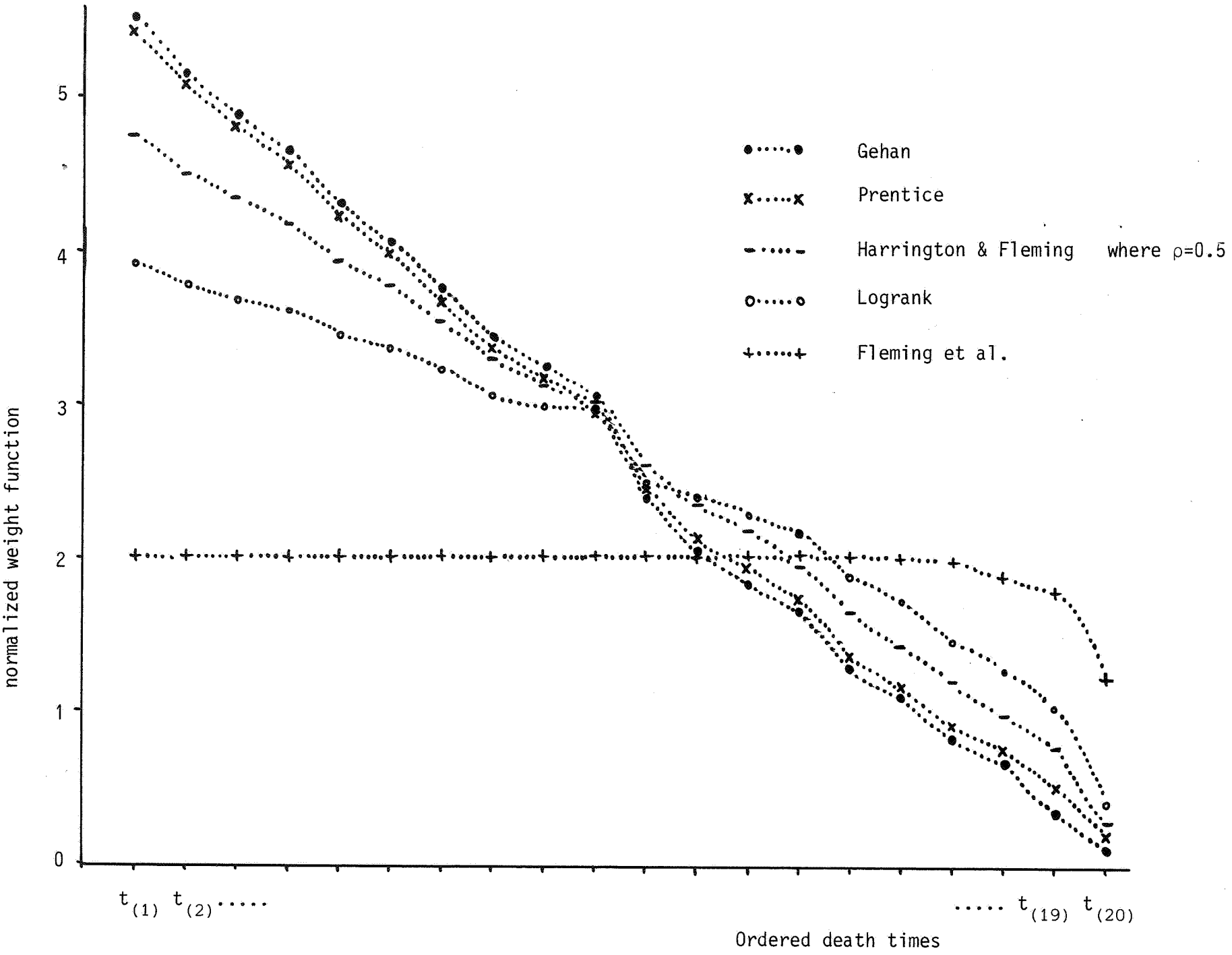

Figure 3: Normalized weight functions of various two-sample tests for the ovarian cancer data.

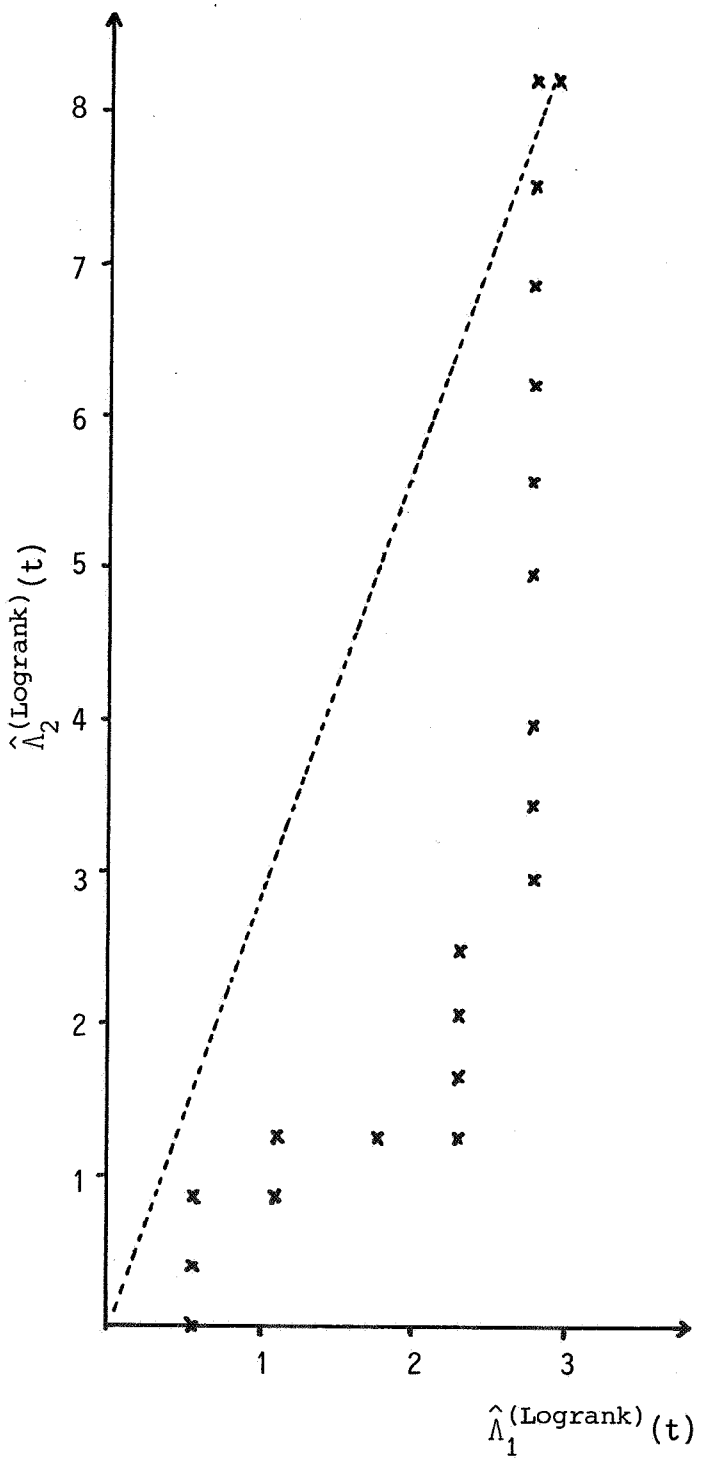

Figure 4: Empirical trend function (xxx) based on the logrank weight function in comparison to the straight line (---) with slope $\hat{\Lambda}_2^{(Logrank)}(\tau) / \hat{\Lambda}_1^{(Logrank)}(\tau)$ (1 ≙ stage II; 2 ≙ stage II a).

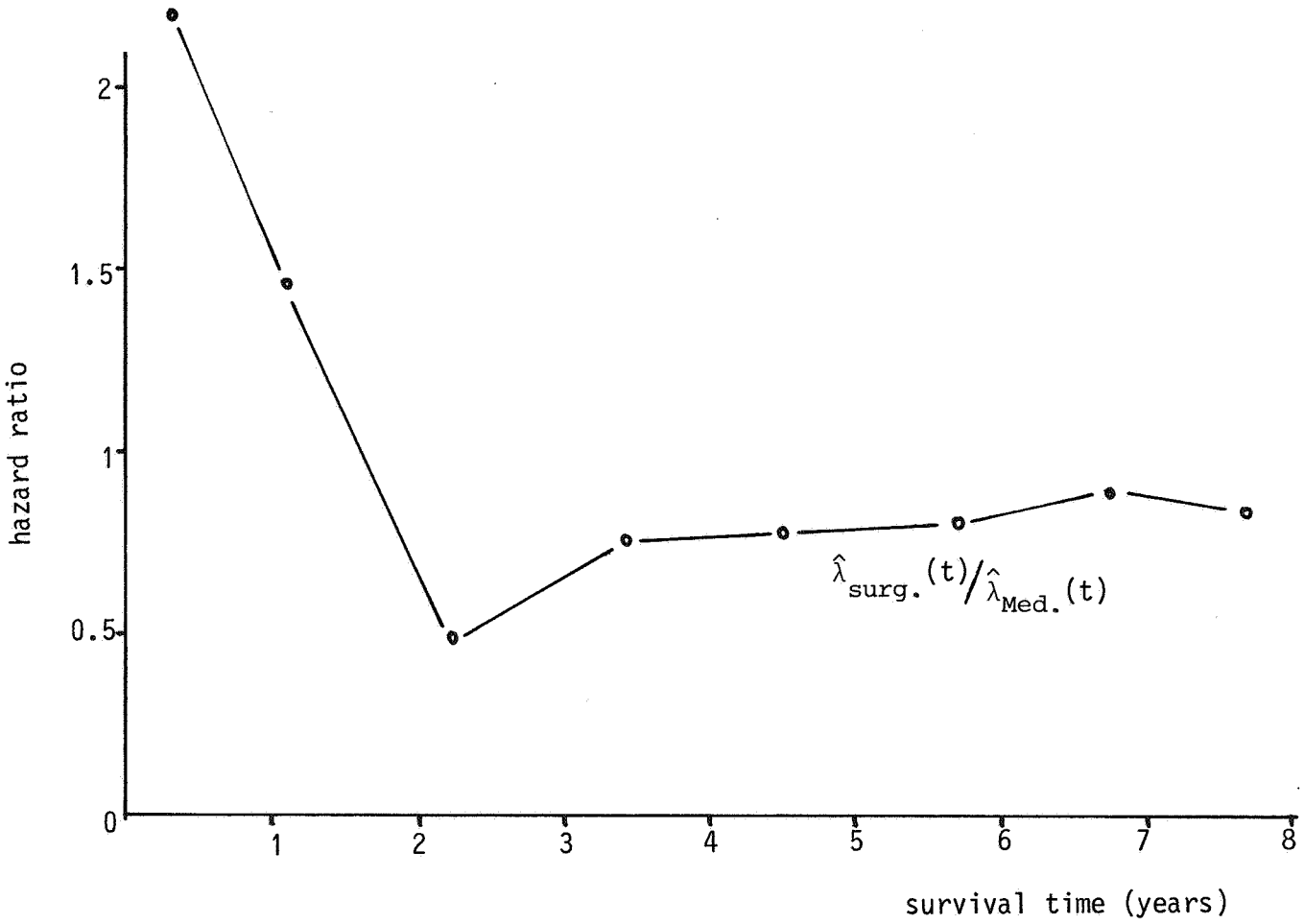

Figure 5: Estimated hazard ratio of coronary artery bypass graft surgery and medical treatment in patients with chronic stable angina (Veteran's Administration trial).

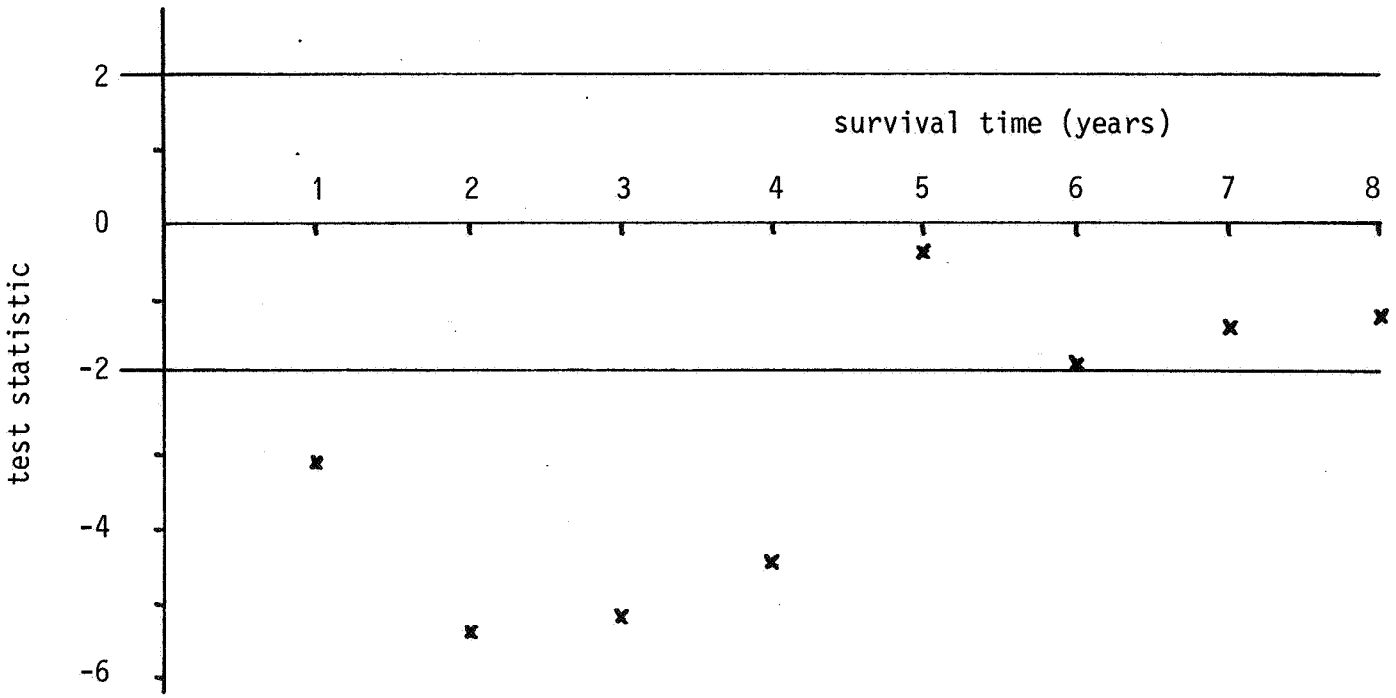

Figure 6: Values of the test statistic $T_{Logrank,Gehan}$ calculated after 1,2,3,...,8 years in the Veteran's Administration trial.

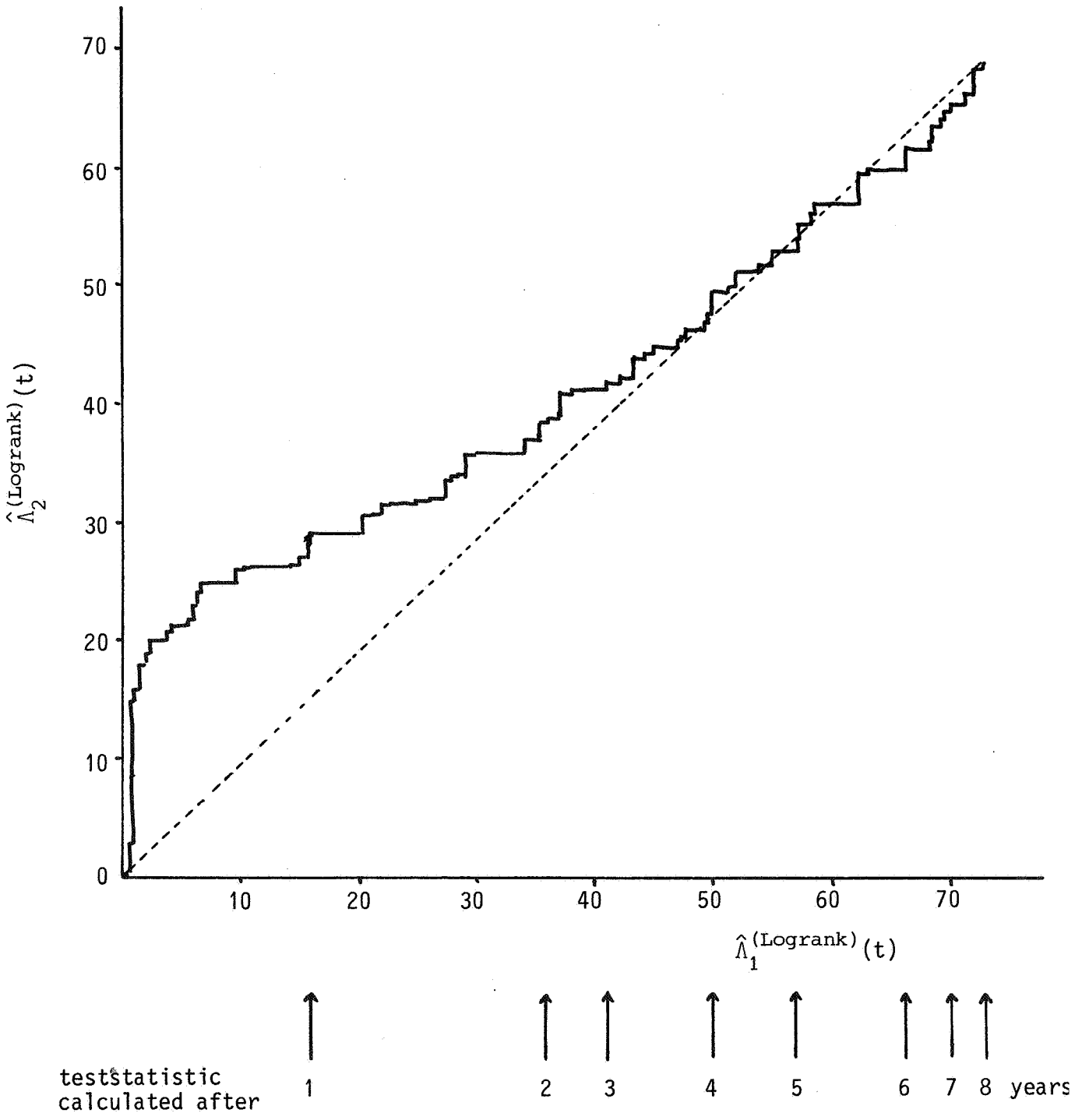

Figure 7: Empirical trend function based on the logrank weight function in the Veteran's Administration example.

REFERENCES

Aalen, O.O. & Hoem, J.M. (1978): Random time changes for multivariate counting processes, Scand. Act. J. 1978, 81- 101.

Andersen, P.K. (1982): Testing goodness-of-fit of Cox's regression model. Biometrics 38, 67-77.

Andersen, P.K. (1983a): Comparing survival distributions via hazard ratio estimates. Scand. J. Statist. 10, 77-85.

Andersen, P.K. (1983b): Testing for proportional hazards. Research Report 81/1, Statistical Research Unit, Univ. of Copenhagen.

Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1982): Linear nonparametric tests for comparison of counting processes, with applications to censored data. Int. Statist. Rev. 50, 219-258. Correction note, Int. Statist. Rev. 52, 225.

Barlow, R.E., Bartholemew, D.J., Bremner, J.M. & Brunk, H.D. (1972): Statistical inference under order restrictions. New York: Wiley.

Barlow, R.E. & Proschan, F. (1969): A note on tests for monotone failure rate based on incomplete data. Ann. Math. Statist. 40, 595-600.

Begun, J.M. & Reid, N. (1983): Estimating the relative risk with censored data. J.Am. Statist. Assoc. 78, 337-341.

Breslow, N. (1984): Comparison of survival curves. In: Cancer clinical trials - methods and practice, 381-406 (eds.: M.E. Buyse, M.J. Staquet & R.J. Sylvester), Oxford University Press, Oxford.

Breslow, N.E., Edler, L. & Berger, J. (1984): A two-sample censored data rank test for acceleration. Biometrics 40, 1049-1062.

Cox, D.R. (1972): Regression models and life tabels (with discussion). J. Roy. Statist. Soc. B 34, 187-220.

Detre, K.M., Hultgren, H. & Takaro, T. (1977): Veterans' Administration Cooperative Study of Surgery for Coronary Arterial Occlusive Disease. III. Methods and baseline characteristics including experience with medical treatment. Am . J. Cardiol. 40, 212-225.

Fleming, T.R., O'Fallon, J.R., O'Brien, P.C. & Harrington, D.P. (1980): Modified Kolmogorov-Smirnov test procedures with application to arbitrarily censored data. Biometrics 36, 607-625.

Freireich, E.J. et al. (1963): The effect of 6-Mercaptpurine on the duration of steroid-induced remissions in acute leucemia. Blood 21, 699-716.

Gehan, E. (1965): A generalized Wilcoxon test for comparing arbitrarily single censored samples. Biometrika 52, 203-223.

Gill, R.D. (1980): Censoring and stochastic integrals. MC Tract 124, Mathematical Centre, Amsterdam.

Gill, R.D. (1984): Understanding Cox's regression model. J. Amer. Statist. Assoc. 79, 441-447.

Gill, R.D. (1985): The total time on test plot and the cumulative total time on test statistic for a counting process. Report MS-R8501, Centrum for Mathematics and Computer Science, Amsterdam.

Kay, R. (1984): Goodness-of-fit methods for the proportional hazards regression model: A review. Rev. Épidém. et Santé Publ. 32, 185-198.

Lee, L. & Pirie, W.R. (1981): A graphical method for comparing trends in series of events. Commun. Statist. - Theor.Meth. A 10, 827-848.

Leurgans, S. (1983): Three classes of censored data rank tests: Strengths and weaknesses under censoring. Biometrika 70, 651-658.

Leurgans, S. (1984): Asymptotic behavior of two-sample rank tests in the presence of random censoring. Ann. Statist. 12, 572-589.

Nagelkerke, N.J.D., Oosting, J. & Hart, A.A.M. (1984): A simple test for goodness-of-fit of Cox's proportional hazards model. Biometrics 40, 483-486.

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

Pocock, S.J., Gore, S.M. & Kerr, G.R. (1982): Long-term survival analysis: The curability of breast cancer. Statistics in Medicine 1, 93-104.

Prentice, R.L. (1978): Linear rank tests with right censored data. Biometrika 63, 291-298.

Prentice, R.L. & Marek, P. (1979): A qualitative discrepancy between censored data rank tests. Biometrics 35, 861-867.

Schoenfeld, D. (1980): Chi-squared goodness-of-fit tests for the proportional hazards regression model. Biometrika 67, 145-153.

Schumacher, M. (1982): Analysis of survival times in nonproportional hazards situations (in German). Habilitationsschrift, University of Heidelberg.

Schumacher, M. (1984): Two-sample tests of Cramér-von Mises- and Kolmogorov-Smirnov-type for randomly censored data. Int. Statist. Rev. 52, 263-281.

Schumacher, M. & Vaeth, M. (1984): On a goodness-of-fit test for the proportional hazards model. EDP in Biology and Medicien 15, 19-23.

Vervaat, W. (1972): Functional central limit theorems for processes with positive drift and their inverses. Zeitschrift f. Wahrscheinlichkeitstheorie verw. G. 23, 245-253.

Wei, L.J. (1984): Testing goodness-of-fit for proportional hazards model with censored observations. J. Amer. Statist. Assoc. 79, 649-652.