AN INTUITIVE NON-RANK-BASED STATISTIC FOR THE TWO-SAMPLE CENSORED-DATA SURVIVAL ANALYSIS PROBLEM

by

Margaret O'Sullivan
Department of Biostatistics
University of North Carolina at Chapel Hill

and

Thomas R. Fleming
Department of Biostatistics
University of Washington

Institute of Statistics Mimeo Series No. 1810

December 1986

An Intuitive Non-Rank-Based Statistic for the Two-Sample Censored-Data Survival Analysis Problem

Margaret O'Sullivan

Department of Biostatistics, 201-H

University of North Carolina

Chapel Hill, NC 27514

and

Thomas R. Fleming,

Department of Biostatistics, SC-32,

University of Washington,

Seattle, WA 98195

Key Words: Two-sample problem; Random Censorship; Stochastic Ordering; Weight Function; Simulations

Summary

A class of statistics based on the integrated weighted difference in Kaplan-Meier estimators is introduced. These statistics are intuitive for and consistent against the alternative of stochastic ordering. The standard weighted Logrank statistics are not always sensitive against this alternative. Qualitative comparisons are made between the weighted Logrank statistics and these SW-statistics. A statement of null asymptotic distribution theory is given and the choice of weight function is discussed in some detail. Results from small sample simulation studies indicate that these statistics compare extremely well with the Logrank procedure and may perform far better than it under the crossing hazards alternative. Extensions to stratification procedures and problems associated with tied data are also discussed.

1. Introduction and Motivation

Consider the classical two-sample survival analysis problem. We assume that the simple random censorship model holds within each of the two groups. The question to be answered is this:

(1.1) 'Is survival in group 1 better than survival in group 2?'

If the survival functions for the two populations cross then without quantifying what we mean by good survival it is unclear what the answer to (1.1) should be. However if the survival functions are ordered uniformly in time then there is a clear answer. Thus any test procedure used to answer (1.1) should be sensitive to the alternative of stochastic ordering. Standard nonparametric procedures such as the Logrank or Wilcoxon tests are not necessarily even consistent against this alternative. In this paper we will develop a simple class of nonparametric procedures which are both intuitive and sensitive for the stochastic ordering alternative. We assume throughout that survival time has a continuous distribution.

Let $\hat{S}_i(\cdot)$ and $\hat{C}_i(\cdot)$ be the product-limit estimators of the (right continuous) survivor functions for the survival and censoring time random variables, namely $S_i(\cdot)$ and $C_i(\cdot)$ respectively, in group i, i = 1, 2. If $\hat{W}(\cdot)$ is a uniformly consistent estimator of the bounded, positive weight function $W(\cdot)$, then define the test statistic

$$(1.2)\quad SW \equiv \sqrt{\frac{n_1 n_2}{n}} \int_0^{T_c} \hat{W}(t)\left(\hat{S}_1(t) - \hat{S}_2(t)\right) dt$$

$$\equiv \sqrt{\frac{n_1 n_2}{n}} \int_0^{T_c} \hat{W}(\hat{S}_1 - \hat{S}_2).$$

Here, n_i is the number of subjects in group i, $n = n_1 + n_2$ and $T_c \equiv \sup \{t: (\hat{S}_1(t) \vee \hat{S}_2(t)) \quad \hat{C}_1(t) \quad \hat{C}_2(t) > 0\}$. SW is the integrated weighted difference between the estimated survival functions. In a sense, it is an 'area' between the curves. The classical nonparametric procedures which have been proposed for this problem are the weighted logrank statistics (Gill 1980). These may be written as

$$L_K \equiv \sqrt{\frac{n_1 n_2}{n}} \int_0^T \hat{K}(t) \left(\frac{dN_1(t)}{Y_1(t)} - \frac{dN_2(t)}{Y_2(t)}\right)$$

$$\equiv \sqrt{\frac{n_1 n_2}{n}} \int_0^T \hat{K} \left(\frac{dN_1}{Y_1} - \frac{dN_2}{Y_2}\right),$$

where $N_i(t)$ is the process which counts the number of observed failures up to and including time t and $Y_i(t)$ is the number of subjects still under observation at time t in group i. The weight function $\hat{K}(\cdot)$ is predictable and $T \equiv \sup\{t: \hat{S}_1(t) \wedge \hat{C}_1(t) \wedge \hat{S}_2(t) \wedge \hat{C}_2(t) > 0\}$.

It can be shown that $\sqrt{\frac{n}{n_1 n_2}}\, SW \xrightarrow{p} \int_0^{\tau_c} W(S_1 - S_2)$ and $\sqrt{\frac{n}{n_1 n_2}}\, L_K \xrightarrow{p} \int_0^{\tau} K(\lambda_1 - \lambda_2)$ where $\lambda_i(\cdot)$ is the hazard function in group i, $K(\cdot)$ is the limiting version of $\hat{K}(\cdot)$ and τ_c and τ are the limiting versions of T_c and T (formal definitions of which are given in the appendix). Thus L_K in essence compares hazard functions. If K is positive L_K will be sensitive to the alternative of ordered hazard functions (which includes the proportional hazards alternative). However if the hazard functions cross, even though stochastic ordering may be maintained, then L_K can have very low power. SW statistics would seem to be more natural than L_K statistics for alterna-

tives which are specified in terms of survival functions. In particular, if $W(\cdot)$ is positive, it is clear that SW will be sensitive to the alternative of stochastic ordering.

Note that L_K uses information on the observation time axis up to T. In fact, [0,T] is the time interval on which the hazard functions for both groups can be estimated since the risk sets in both groups are positive on [0,T]. The SW statistic on the other hand can use information accumulated on $[0,T_c]$ It can be shown that the survival functions $S_1(\cdot)$ and $S_2(\cdot)$ are well estimated in both groups on $[0,T_c]$ (O'Sullivan 1986) and thus $[0,T_c]$ is the natural interval on which to compare the survival functions. Note that, in general $T \leq T_c$ and therefore SW will include information from more of the observation time axis than L_K in some cases.

In addition to their more intuitive appeal than the L_K statistics for detection of the alternative of stochastic ordering, SW statistics are basically different from L_K statistics in a second property. L_K statistics are generalized rank statistics, placing mass only at observed death times. SW statistics on the other hand are not rank statistics in general since the integration in (1.2) is with respect to Lebesgue measure. The generalized rank statistics are certainly robust. They are also (with the usual choices for $\hat{K}$) invariant under monotone increasing transformations of the data. However, this is not necessarily a desirable feature, particularly if it is important that large differences in the magnitude of survival time are detected with higher probability than small differences. For example, consider the two situations depicted in Figure 1. Certainly, there is much

more evidence of a large gain in survival time in group 1 over group 2 in configuration II than there is in configuration I. Yet, an L_K-statistic will yield the same value for both I and II. In contrast, for reasonable choices of $W(\cdot)$, the SW statistic calculated will be much larger in II than in I. We contend that SW statistics are more suitable than the generalized linear rank statistics for detecting large differences in the magnitude of survival time. This will become obvious in section 3 which contains small sample simulation results.

A statement of asymptotic distribution theory results for our new class of non rank censored survival data statistics is contained in section 2. In section 3 the choice of weight function is discussed at some length. A variety of small sample simulation results are presented in section 4. Finally extensions to stratification procedures and problems associated with tied data are discussed in section 5.

2. Asymptotic Distribution Theory

The detailed development of asymptotic distribution theory for SW is given in O'Sullivan and Fleming (1986). Here, we will merely state some of the results under the null hypothesis so that asymptotically valid test procedures can be formulated. Results on consistency and efficiency also appear in O'Sullivan and Fleming (1986).

In this section assume that H_o: $S_1 = S_2$ holds and denote the common survival function by S. Note that, under the null, $\tau=\tau_c$. We also need a definition.

Definition

For two sequences of stochastic processes X^n and Y^n, $X^n = O(Y^n)$ if $\exists$ a constant Γ such that given $\epsilon > 0$, $\exists\ n_\epsilon$ with

$$P[|X^n(t)| \leq \Gamma|Y^n(t)|\ \forall\ t\geq 0,\ \forall\ n \geq n_\epsilon] \geq 1-\epsilon.$$

Theorem 2.2

Suppose the following assumptions hold (for i = 1,2).

(A1) $\lim_{n\to\infty} \frac{n_i}{n} = \rho_i > 0$,

(A2) $\tau < \infty$,

(A3) $\sup_{t\in(T_c,\tau)} |\hat{W}(t) - W(t)| \xrightarrow{P} 0$,

(A4) $W(\cdot) = O((C_i^-(\cdot\wedge\tau_i))^{\frac{1}{2}+\delta})$

and $\hat{W}(\cdot) = O((\hat{C}_i^-(\cdot))^{\frac{1}{2}+\delta})$

for some $\delta > 0$,

then

$$SW \xrightarrow{d} N(0, \sigma^2)$$

where $\sigma^2 = \sum_{i=1}^{2} - \rho_{3-i}\int_0^\tau \frac{(\int_u^\tau WS)^2}{S^2C_i^-}\,dS$

Lemma 2.3

Let

$$\hat{\sigma}^2_{up} \equiv - \sum_{i=1}^{2} \hat{\rho}_{3-i}\,\frac{n_i}{n_i-1}\int_0^{T_c} \frac{(\int_u^{T_c}\hat{W}\hat{S}_i)^2}{\hat{S}_i\hat{C}_i^-\hat{S}_i^-}\,d\hat{S}_i ,$$

and

$$\hat{\sigma}^2_p = -\sum_{i=1}^{2} \hat{\rho}_{3-i} \frac{n}{n-1} \int_0^{T_c} \frac{\left(\int_u^{T_c} \hat{W} \hat{S}_p\right)^2}{\hat{S}_p \hat{C}_i \hat{S}_p} d\hat{S}_p \ ,$$

where $\hat{S}_p$ is the Kaplan-Meier estimator from the pooled sample and $\hat{\rho}_i = \frac{n_i}{n}$. Then

$$\hat{\sigma}^2_{up} \xrightarrow{p} \sigma^2$$

and

$$\hat{\sigma}^2_p \xrightarrow{p} \sigma^2$$

under the assumptions of Theorem 2.2.

Asymptotically valid test procedures can be formulated by a comparison of $SW/\sqrt{\hat{\sigma}^2_{up}}$ or $SW/\sqrt{\hat{\sigma}^2_p}$ with the standard normal quantiles. The performance of such procedures in small samples is explored in section 4.

3. The Choice of Weight Function

The choice of weight function will generally depend on the application in mind. Three basic considerations are involved in the choice. These are stability, interpretability and efficiency. We will deal with each of these considerations here.

The test procedure should be (asymptotically) valid for all possible underlying choices of survival and censoring configurations. In this sense the test procedure is nonparametric, and the weight function should be chosen to guarantee this. In particular to achieve stability with finite variance σ^2,

the constraint $\hat{W}(\cdot) = o\left((\hat{C}_i^-(\cdot))^{\frac{1}{2}+\delta}\right)$ for some $\delta > 0$ should be satisfied. Note that if one uses the statistic $SW(t_o) \equiv \sqrt{\frac{n_1 n_2}{n}} \int_o^{t_o} \hat{W}(\hat{S}_1 - \hat{S}_2)$ where t_o is a fixed time-point with $t_o < \tau$, then this constraint on the weight function could be eliminated. However $SW(t_o)$ can still have very high variance and hence very low power for particular choices of censoring distributions regardless of the differences in survival functions which may occur. Hence, use of $SW(t_o)$ in practice is not recommended over imposition of the constraint $\hat{W}(\cdot) = o\left((\hat{C}_i^-(\cdot))^{\frac{1}{2}+\delta}\right)$.

The weight function should yield the statistic SW interpretable. If the weight function is unity then $\sqrt{\frac{n}{n_1 n_2}}$ SW estimates $\mu_1 - \mu_2$, the difference in mean survival in $[0,\tau_c]$. Although $\mu_1 - \mu_2$ is interpretable, $W(\cdot) \equiv 1$ does not satisfy the constraint (A4). Indeed in order that the stability constraint be satisfied in general, it is necessary that the weight function be a function of the estimated censoring distributions. A weight function which reduces to unity in uncensored data and which does satisfy (A4) is

$$\hat{W}_c(\cdot) \equiv \left(\frac{\hat{C}_1^-\hat{C}_2^-}{\hat{\rho}_1\hat{C}_1^- + \hat{\rho}_2\hat{C}_2^-}\right) \quad (\cdot) \ .$$

In data which is heavily censored so that (A4) is not satisfied by $\hat{W}(\cdot) = 1$ (i.e. if $C_i^-(\tau_i) = 0$ i=1 or 2) then $\mu_1 - \mu_2$ cannot be estimated with precision due to instability introduced by the censoring. Under stochastic ordering $\sqrt{\frac{n}{n_1 n_2}}\ SW_c \equiv \int_o^{T_c} \hat{W}_c(\hat{S}_1 - \hat{S}_2)$ estimates a lower bound on $\mu_1 - \mu_2$ in censored data with precision of the order of $1/\sqrt{n}$. . Note that if the

censoring distributions are equal then $\sqrt{\frac{n}{n_1 n_2}} SW_c$ in fact estimates the difference in mean observation times.

The weight function need not reduce to unity in uncensored data to maintain interpretability of SW there. In industrial and marketing settings for example, $\int_0^{\tau_c} W(S_1 - S_2)$ might give a measure of the cost difference in one method of production over another.

Example 3.1

An experiment is carried out in a wood-processing plant to compare two chemical treatments of wood planks. The purpose of the treatments is to enhance the strength of the planks. Each plank is randomized to one of the two treatments and tested for strength by the application of a gradually increasing load to the plank until it breaks. The "load until breakage" is the survival time random variable. If $V(u) = P[\text{a plank is used at load} \leq u \text{ in practice}]$ then the expected proportion of breakages in practice for planks given treatment i is $\int_0^{\infty}(1 - S_i(u))dV(u) = \int_0^{\infty} v(u)(1-S_i(u))du$ where $v(u) = \frac{d}{du}V(u)$. Hence in terms of real cost the natural measure on which to base the comparison of the two treatments is $\int_0^{\infty} v(u)(S_1-S_2)(u)du$.

Censoring will occur if the load is not increased up to the breaking point for some planks. In this case a procedure to test for the equality of treatments can be based on $\int_0^{T_c} \hat{W}_c(u)v(u)(\hat{S}_1(u) - \hat{S}_2(u))du$. The test is sensitive to the alternative of stochastic ordering and $\sqrt{\frac{n}{n_1 n_2}}\ SW(T_c)$ estimates a lower bound on $\int_0^{\tau_c} v(u)(S_1 - S_2)(u)du$ in this case.

In a clinical trial W might involve some notion of quality of life.

Example 3.2

In the treatment of cancer by chemotherapy if the treatments are particularly aggressive then interest may lie solely in long-term survival. Survival differences which occur during the course of the aggressive treatments are ignored. Thus if treatments are administered over $[0,t_o]$ the choice $W(t)=0$ $t<t_0$, $W(t)=i$ $t\geq t_0$, yields an estimate $\sqrt{\frac{n}{n_1 n_2}}SW(T_c)$ of the difference in mean "livable lifetime" over $[0,\tau_c]$ in uncensored data. In censored data and under stochastic ordering past t_0, $W(t) \equiv W_c(t)I\{t\geq t_0\}$ yields an estimate of a lower bound on this difference. Note that statistics based on hazard functions have no capacity for dealing with the quality of life aspect of such a clinical trial. The pattern of deaths on the two treatment arms past t_0 may in no way reflect differences in survival probabilities past t_0.

In the examples above the quantity of interest is well estimated in uncensored data. Censoring however introduces difficulties into the estimation of quantities based on real time. Stable estimation requires down-weighting of information over periods of heavy censoring. Indeed intuitively it is clear that censored data is intrinsically unsuitable for estimation of real-time parameters. However, realizing the limitations of the data, in the discussion above test procedures were developed which, by virtue of the particular choice of weight function, are sensitive to the alternative of most interest. In addition an estimate of a lower bound on the quantity of interest (which is an interpretable quantity) is obtained assuming the alternative of interest is true. These types of techniques have not been considered in

survival analysis before.

Finally, the weight function should render the statistic efficient in that alternatives of most interest are detected with high probability. Suppose an alternative of interest can be written as H_1: $S_1(t) - S_2(t) \simeq D(t)$. Although the most powerful test for such an alternative in small samples cannot be determined, the locally most powerful test under a sequence of local alternatives given by H_1^n: $\sqrt{\frac{n_1 n_2}{n}}(S_1^n - S_2^n)(t) \longrightarrow D(t)$ can often be determined. Indeed if $D(\cdot)$ is bounded and the convergence is uniform then, under some mild regularity conditions given in Lemma 4.2 of O'Sullivan and Fleming (1986) for a given limiting configuration of continuous survival and censoring distributions (S^o, C_1^o, C_2^o) the optimal weight function is given by

$$W_{opt}(v) \propto \frac{1}{S^o(v)} \frac{d}{dv} \left\{ (S^o(v))^2 \left(\frac{C_1^o C_2^o}{\rho_1 C_1^o + \rho_2 C_2^o} (v) \right) \frac{\frac{d}{dv} \frac{D(v)}{S^o(v)}}{\frac{d}{dv} S^o(v)} \right\}$$

As noted in O'Sullivan and Fleming (1986) this deterministic weight function does not in general yield a non-parametric test. Although SW_{opt} is most efficient for that particular choice of survival and censoring distributions, it may not yield a weakly convergent test statistic under the null for a different configuration. Thus although W_{opt} is of some academic interest and indeed is informative in accessing the behaviour of particular SW statistics in given situations (O'Sullivan (1986)), it is not truly of practical interest.

For the simulation studies to be presented in the next section we chose

the weight functions $\hat{W}_c$ and $\hat{W}_{\underline{V}c} \equiv (\hat{W}_c)^{\frac{1}{2}}$. Since the only interest in these experiments was the detection of the general alternative of stochastic ordering these simple, intuitive generalized z-tests were appropriate. $\hat{W}_c$ satisfies the regularity conditions of Theorem 2.2. Indeed $(\hat{W}_c)^{\frac{1}{2}+\delta}$, satisfies them for any $\delta>0$. Thus we investigated the performance of the limiting version as δ tends to zero, namely $\hat{W}_{\underline{V}c}$, in the simulation studies also.

4. Small Sample Simulation Results

Let $SW_c \equiv \underline{\sqrt{\frac{n_1 n_2}{n}}} \int_o^{T_c} \hat{W}_c(\hat{S}_1 - \hat{S}_2)$ and $SW_{\underline{V}c} \equiv \underline{\sqrt{\frac{n_1 n_2}{n}}} \int_o^{T_c} \hat{W}_{\underline{V}c}(\hat{S}_1 - \hat{S}_2)$. The notation $SW^p{}_c$ and $SW^p{}_{\underline{V}c}$ will be used to denote SW_c and $SW_{\underline{V}c}$ respectively standardized by $\underline{\sqrt{\hat{\sigma}^2_p}}$. Similarily $SW^{up}_c \equiv SW_c / \sqrt{\hat{\sigma}^2_{up}}$ and $SW^{up}_{\underline{V}c} \equiv SW_{\underline{V}c} / \sqrt{\hat{\sigma}^2_{up}}$. In the simulation studies we performed, the behaviours of these statistics were compared with that of the Logrank statistic (denoted by Lgk). The Logrank is generally regarded as the standard test procedure for the two-sample censored data survival analysis problem. Comparisons were also made with the Prentice-generalized Wilcoxon statistic (Wlx) which is another widely used statistic. All test procedures presented here are one-sided with the alternative of interest being H_1: $S_1(\cdot) \geq S_2(\cdot)$, $S_1 \neq S_2$. The simulations were performed on an IBM-AT personal computer, using the APL programming language and it's inherent random number generator.

4.1 Size Properties

All survival and censoring distributions used in this study were continuous. A variety of underlying survival distributions and a variety of censoring distributions for each survival distribution were used. This differs from the usual size simulation study design for generalized rank statistics. For such studies it is enough to consider a single continuous survival distribution and a range of censoring distributions due to the invariance of rank statistics under monotone increasing transformations of the data. SW statistics do not share this invariance property and hence a more involved study was necessary.

The survival distributions chosen were Weibull with scale parameters b=1 and shape parameters a=.5,1,2,3 where $S_{b,a}(t) = e^{-(t/b)^a}$, $t \geq 0$, denotes the Weibull survival function. This group of survival functions is diverse in terms of skewness and tailweight, factors which might effect the empirical sizes of the tests under study. The censoring distributions chosen were equal and uniform in the studies we present here. The expected proportion censored under the various survival and censoring configurations and the simulation results under the null hypothesis are displayed in Table 1.

In almost all cases studied SW_c and $SW_{\underline{V}c}$ provide equally acceptable empirical significance levels, with the test based on $SW_{\underline{V}c}$ being slightly more anticonservative than SW_c in general. From a practical point of view we can conclude that $SW_{\underline{V}c}$ is equally as acceptable to use in small samples as SW_c despite the lack of asymptotic distribution theory for it. Indeed the

simulation results suggest that the asymptotic distribution theory for $SW_{\underline{V}c}$ may in fact be valid though assumption (A4) does not hold.

The choice of variance estimator has a large influence on the null behaviour of the SW statistic. Tests based on the unpooled variance estimator $\hat{\sigma}_{up}^2$ are much more anticonservative than tests based on the pooled variance estimator. Indeed by the usual variance decomposition we know this to be true in uncensored data. From Table 1 it is clear that use of the unpooled variance estimator often leads to an unacceptably anticonservative test especially at lower p-values. In unequally censored data the situation worsens considerably for $\hat{\sigma}_{up}^2$ and yet the $\hat{\sigma}_{p}^2$ performs very well (O'Sullivan 1986). Thus we recommend use of the pooled variance estimator in practice.

The empirical sizes of SW_c^p and $SW_{\underline{V}c}^p$ are both very close to the nominal level across a broad range of situations. Only in uncensored heavy tailed survival distributions do the empirical levels deviate substantially from the nominal levels and then only at the lower significance level of .01 rather than at the .05 level. Because we are essentially dealing with the z-test in uncensored data it is not surprising that the tests are conservative in uncensored heavy tailed distributions. (Benjamini (1983) has shown that the conservativism occurs primarily at low p-values). Indeed to protect against such situations, in classical statistics one might 'trim the data', using trimmed sample means for comparison rather than true sample means. That is to say, we would artificially censor the data to increase robustness of the test procedure! The natural censoring seen in survival data should yield the SW^p statistics robust in real applications.

4.2 Power Properties

We only present results for SW^p_c and $SW^p_{\underline{V}c}$ in the power studies due to the good size properties of these statistics. A much more detailed study of various aspects of the small sample behaviour of SW statistics, including the power properties of SW^{up}_c and $SW^{up}_{\underline{V}c}$, is given in O'Sullivan (1986). The interested reader is referred to that source.

(i) The Proportional Hazards Alternative

Due to the current popularity of the proportional hazards model and the availability of the locally most powerful test (the Logrank) for this alternative under equal censoring we investigated the performance of SW statistics under this alternative. Again the survival distributions chosen were Weibull with the shape parameter for S_1 and S_2 being the same within a particular configuration. The scale parameters were chosen so that the Logrank test had a reasonable probability of detecting the constant hazard ratio, γ, at the sample sizes and significance levels chosen and for a variety of censoring distributions.

The simulation results indicate that $SW^p_{\underline{V}c}$ is remarkably almost as powerful as the logrank in detecting the Weibull proportional hazards alternative under equal uniform censoring distributions. The empirical relative efficiency is denoted by $EE(SW^p_{\underline{V}c}, Lgk)$ and is defined as $P(SW^p_{\underline{V}c})/P(LgK)$ where P(T) is the number of times the null was rejected using the test statistic T. This was at least 90% in almost all cases

examined. In uncensored data SW^p also seems to be efficient except for the very heavy tailed survival distributions with a = $\frac{1}{2}$, and then most noticably at the lower significance level. This is not surprising due to the conservativism of SW in this situation under the null hypothesis. $SW^p_{\underline{V}c}$ has slightly higher power than SW^p_c in censored data for proportional hazards alternatives. Indeed, this is to be expected since large differences in the survival functions occur later rather than earlier under the configurations chosen and $SW^p_{\underline{V}c}$ places relatively more weight on later survival differences than SW^p_c does. The difference is however very small.

An in depth study of the performance of the Logrank and SW statistics under unequal censoring patterns is given in O'Sullivan (1986). The study indicates that the power of $SW^p_{\underline{V}c}$ is very close to that of the Logrank under unequal censoring patterns except in cases where their true sizes differed under the null under the same censoring patterns and for the same Weibull shape parameter. This occurs primarily at the lower α-level of .01 where the logrank is known to be a very anticonservative test (Breslow et al (1984)). Indeed in such cases it is questionable whether the logrank is even an acceptable test in terms of size. Certainly $SW^p_{\underline{V}c}$ is far more stable than it under unequal censoring patterns under the null. However, because of this, $SW^p_{\underline{V}c}$ can be far less powerful than the Logrank in some cases under the alternative.

(ii) Periods of Equality of Hazard Functions.

The generalized linear rank statistics are based on differences in estimated hazard functions. The power of the statistic is essentially governed

by the factor $\int_0^{\tau_c} K(\lambda_1-\lambda_2)$. Thus periods over which survival differences are large but over which hazard functions are equal do not contribute to the power of the L_K test procedure. Indeed such periods decrease the power of the statistic if hazards are non-zero due to the increased variability in the numerator. Consider the piecewise exponential configurations of Figure 2 and the corresponding simulation results in Table 3. In configuration I the Logrank and SW_c^p procedures are equally sensitive to differences in survi over (0,.5). The zero hazard difference over (.5,1) however yields more sensitive to the entire alternative of configuration I for the reas cited above. The introduction of a period of zero hazard (on a lag-time) i configuration II increases the power of SW_c^p but does not effect the power of the logrank. This is because the logrank places no mass during periods of zero hazard, since no deaths occur there, yet the magnitude of the difference in survival relative to the overall variability increases from configurations I to II. This is apparent by looking at the configurations in Figure 2.

SW_c^p is scale invariant, a desirable property in that the power of the test does not depend on the units chosen to measure time! The effect of scale invariance is also seen in Table 3. There is no significant gain in power in configurations III and IV over configuration II. It is also apparent in Figure 2 that there is no significant increase in the scaled observed mean difference in survival time in configurations III and IV over II.

(iii) Versatility and Crossing Hazards

A final set of simulations were performed to investigate whether or not

SW^p_c and SW^p_{Vc} are versatile test statistics in the sense that they can detect a range of stochastic ordering alternatives. The configurations chosen are displayed in Figure 3.

The underlying survival functions are piecewise exponential in each case. The configurations were chosen so that they would include a range of differences that might be seen in practice. For example, in a clinical trial to compare an experimental treatment with a placebo the correspondence between the configurations of Figure 3 and the treatment effect is as follows. In I the treatment decreases the constant hazard rate by a factor which is constant over time. In II and III the treatment decreases the hazard rate only over a particular period of time, the time period being early in II and late in III. The hazard functions cross in VI and V. The effect of treatment is to decrease the hazard rate initially but the treated population has a higher hazard rate than the untreated population later on so that long-term survival probabilities are the same. Configuration V includes an intermediary time period during which the hazard rates are small and equal in the two groups. Configurations I and II have essentially been discussed above (i) and (ii) respectively. In III no difference can be detected by any of the statistics in censored data. In uncensored data the Wilcoxon performs very poorly because it is insensitive to differences which only occur late in time. It places most weight on early hazard differences. Both the Logrank and SW^p_c statistics are equally powerful in detecting these later differences.

The poor performance of the Wilcoxon relative to the Logrank under an alternative like III is one reason for the high relative popularity of the Logrank over the Wilcoxon. In practice detection of large long term differences is often very important and the Wilcoxon may not detect this as was seen in our simulations. The Logrank performs poorly in IV since this is a case of crossing hazards. The Wilcoxon performs much better since it places most weight on the early positive hazard difference $\lambda_2 - \lambda_1$ and little weight on the later negative hazard difference. In censored data SW^p_c performs very well but loses it's superiority over the logrank in uncensored data. This is reasonable since the observed difference in survival relative to overall variability under uniform (0,2) censoring is quite large, but is decreased substantially when censoring is removed. The introduction of a lag-time in V however increases the efficiency of SW^p relative to the Wilcoxon in uncensored data under crossing hazards.

In conclusion it is clear that the power of SW^p_c or $SW^p_{\underline{V}c}$ is determined by the magnitude of the difference in observed survival time scaled by the overall variability, rather than by the difference in hazard functions as is the case for the generalized linear rank statistics. Across a broad range of alternatives SW^p_c (or $SW^p_{\underline{V}c}$) is a good competitor to the standard Logrank test even under the proportional hazards alternative. It can perform substantially better than the Logrank when the hazard functions cross.

5. Extensions and Discussion

5.1 Stratification.

A stratified version of the $SW_c(T_c)$ test is easily computed as follows. Suppose there are k_0 levels of a stratification variable z and that the alternative of interest is stochastic ordering in the same direction within each level of z, i.e. $H_1: S_{1,k}(\cdot) \geq S_{2,k}(\cdot)$ $k=1,\ldots,k_0$ with $S_{1,k}(\cdot) \neq S_{2,k}(\cdot)$ for some k. An appropriate stratified SW-statistic is

$$SW^S \equiv \frac{\sum_{k=1}^{k_0} \sqrt{\frac{n_{1,k}n_{2,k}}{n_k}} \int_0^{T_{c,k}} \hat{W}_k(\hat{S}_{1,k}-\hat{S}_{2,k})du}{\sqrt{\sum_{k=1}^{k_0} \sum_{i=1}^{2} \hat{\rho}_{3-i,k} \frac{n_k}{n_k-1} \int_o^{T_{c,k}} \frac{\left(\int_v^{T_{c,k}} \hat{W}_k \hat{S}_{p,k}\right)^2}{\hat{S}_{p,k}\hat{C}_{i,k}^{-}\hat{S}_{p,k}^{-}} d\hat{S}_{p,k}}}$$

where the subscript k denotes the level of z and we use the pooled variance estimator although an unpooled variance estimator may also be used.

This statistic will have a standard normal distribution asymptotically under the null hypothesis if the regulatory conditions of Theorem 2.2 are satisfied for the two groups within each level of z. In particular it is important that $\hat{W}_k(\cdot) = 0\left((\hat{C}_{i,k}^{-}(\cdot))^{\frac{1}{2}+\delta}\right)$ $i = 1,2$ for $k=1,\ldots,k_0$. The strata may be weighted differently in the statistic to enhance interpretability or to increase efficiency against particular alternatives.

If the less specific two-sided alternative is of interest, $H_1: S_{i,k}(\cdot) \geq S_{3-i,k}(\cdot)$ $i=1$ or $2,k=1,\ldots,k_0$, and $S_{1,k_1} \neq S_{2,k_1}$ for some k_1 then a chi-squared test can be formulated as

$$\sum_{k=1}^{k_0} \frac{n_{1,k}n_{2,k}}{n_k} \frac{\left(\int_0^{T_{c,k}} \hat{W}_k(\hat{S}_{1,k} - \hat{S}_{2,k})du\right)^2}{\hat{\sigma}^2_{p,k}(T_{c,k})}$$

This has a $\chi^2_{k_0}$ distribution under the null.

5.3 Discussion

The class of SW statistics presented here are both intuitive and sensitive for the alternative of stochastic ordering. The statistics are essentially different from the popular generalized linear rank statistics and are much closer to the z-test in philosophy. Other generalized rank procedures which have been proposed for this problem include the generalized Smirnov test (Fleming et al 1980), the Supremum tests (Fleming et al 1986), the generalized Cramer von Mises and Smirnov tests (Schumacher 1984) and the Logrank-acceleration test of Breslow et al (1984). Generally these statistics have been developed as more versatile procedures than the linear rank statistics. A comparison of SW statistics with these procedures should be made.

In practice the generalized linear rank statistics often perform well, and reporting the p-value of the Logrank or Wilcoxon is the standard procedure at present. In order to facilitate use of the SW statistic without abandoning the standard procedure one might base a statistical test on a cluster of statistics, namely an SW statistic with a weighted logrank statistic. Such a procedure is also more versatile than use of one statistic alone. This is similar to the logrank-acceleration test of Breslow et al (1984) and to a

procedure suggested by Fleming et al (1986). Since the Wilcoxon and SW statistics seem to detect quite different alternatives a very versatile test might be obtained with this cluster. It is shown in O'Sullivan (1986) that the cluster has an asymptotically joint normal distribution and consistent estimators of the variance covariance matrix are easily calculated.

A basic assumption made in particular for the development of asymptotic distribution theory is that $S_i(\cdot)$ is continuous for i=1 and 2. Although survival time may in truth have a continuous distribution, in practice data is always recorded in discrete units. Hence ties will often occur in real data. If we can assume that the recording unit is small, that in truth censored observations occur after survival time observations at tied data points, and that the underlying survival time distributions are continuous, then it is shown in O'Sullivan (1986) that the statistic calculated from it's definition (with the ties incorporated in the usual definition of the Kaplan-Meier estimators etc) is a close approximation to that calculated had the true survival observation times been recorded. Since SW-statistics are real time statistics this is to be expected. However if the underlying survival distribution is truly discontinuous then further theoretical and simulation work will be necessary to determine valid test procedures based on SW-statistics.

The two-sample procedure can be generalized in various ways to more than two samples. Both trend statistics and omnibus k-sample test statistics can be formulated in terms of the estimated survival functions. These procedures along with details of procedures based on clusters will be included in a later publication.

Acknowledgements: We are grateful to Diane Lloyd for her assistance in the preparation of this manuscript.

References

Benjamini, Y. (1983). Is the t-test really conservative when the parent distribution is long-tailed? *JASA* **78**: 845-54.

Breslow, N.E. , Edler, L. and Berger, J. (1984). A two-sample censored data rank test for acceleration. *Biometrics* **40**: 1049-62.

Fleming, T.R., O'Fallon, J.R., O'Brien, P.C. and Harrington, D.P. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right censored data. *Biometrics* **36**: 607-25.

Fleming, T.R., Harrington, D.P. and O'Sullivan, M. (1986). Supremum versions of the Logrank and Wilcoxon statistics. *JASA* (in press).

Gill, R.D. (1980). *Censoring and Stochastic Integrals.* Mathematical Center Tracts 124. Mathematisch Centrum, Amsterdam.

O'Sullivan, M. (1986). *A New Class of Statistics for the Two-Sample Survival Analysis Problem.* Ph.D. thesis, University of Washington.

O'Sullivan, M. and Fleming, T.R. (1986) Statistics for the two-sample survival analysis problem based on product limit estimators of the survival functions. *University of North Carolina, Center for Stochastic Processes, Tech. Report No. 163.*

Schumacher, M. (1984). Two-sample tests of Cramer-von Mises and Kolmogorov-Smirnov type for randomly censored data. *International Statistical Review* **52**: 263-81.

Appendix

Definition

$$\tau_i \equiv \sup\{t:\ S_i(t)\ \ C_i(t) > 0\}.\ i=1,2.$$

$$\tau \equiv \tau_1\ \ \tau_2$$

$$\tau_c \equiv \begin{cases} \tau_{3-i} & \text{if } S_i(\tau_i) = 0 \text{ and } C_i^-(\tau_i) > 0\ ,\ \tau_i < \tau_{3-i} \\ \tau_i & \text{otherwise} \end{cases}$$

Algebraic Forms for the Configurations of Figure 4.

$$S_i(v) = \exp\left(-\int_0^v \lambda_i(t)dt\right)$$

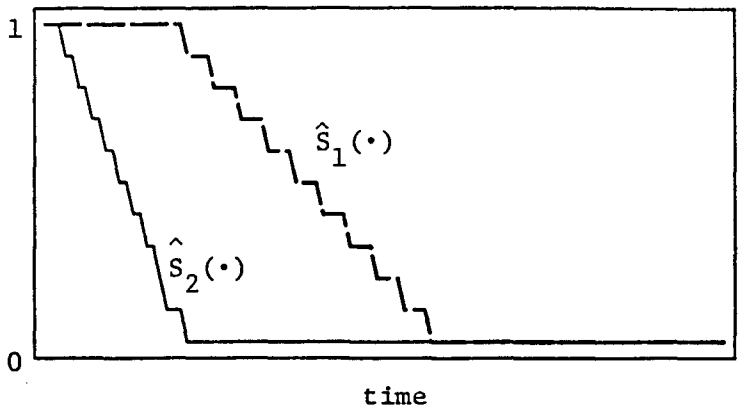
I
$$\lambda_1(t) = .5 \qquad t\geq 0$$
$$\lambda_2(t) = .8 \qquad t\geq 0$$

II
$$\lambda_1(t) = .25I\{t \leq .75\} + .5I\{t > .75\}$$
$$\lambda_2(t) = .75I\{t \leq .75\} + .5I\{t > .75\}$$

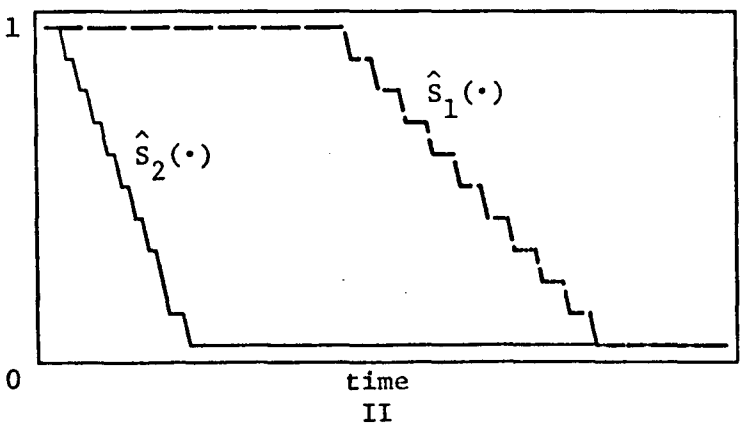
III
$$\lambda_1(t) = .5 \quad t \geq 0$$
$$\lambda_2(t) = .5I\{t \leq .5\} + 1I\{t > .5\}$$

IV
$$\lambda_1(t) = .5I\{t \leq .75\} + 1.5\{.75 < t \leq 1.5\} + I\{t > 1.5\}$$
$$\lambda_2(t) = 1.5I\{t \leq .75\} + .5I\{.75 < t \leq 1.5\} + I\{t > 1.5\}$$

V
$$\lambda_1(t) = .5I\{t \leq .5\} + .1I\{.5 < t \leq 1.25\} + 1.5I\{1.25 < t \leq 1.75\} + I\{t > 1.75\}$$
$$\lambda_2(t) = 1.5I\{t \leq .5\} + .1I\{.5 < t \leq 1.25\} + .5I\{1.25 < t \leq 1.75\} + I\{t > 1.75\}$$



I



II

Figure 1

Configurations exhibiting a small (I) and larger (II) difference in the magnitude of survival time.

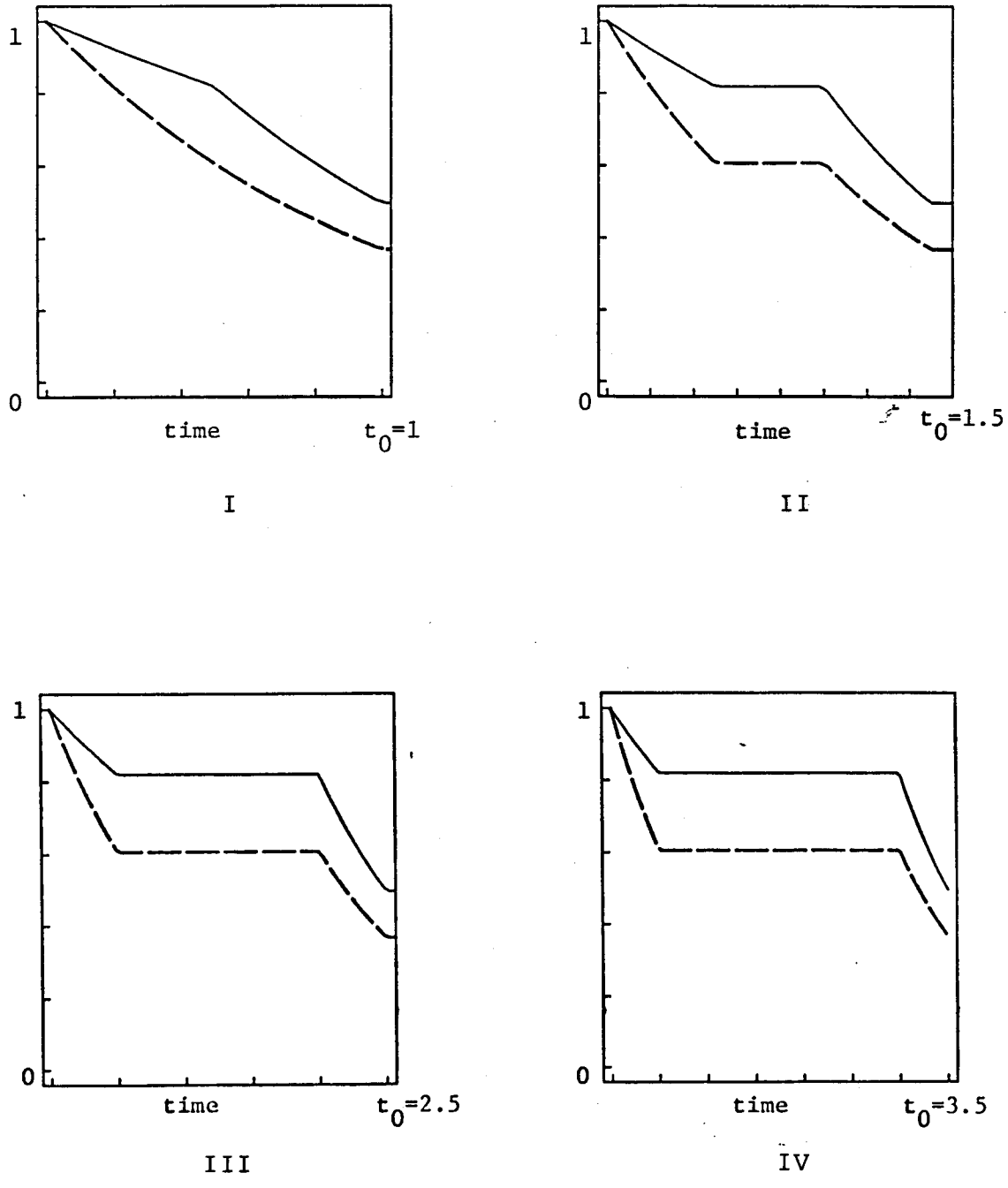


Figure 2

Survival configurations with hazard rates which are proportional early in time $(0,.5)$ $(\lambda_1 = .4, \lambda_2 = 1)$ and equal late in time $(t_0-.5, t_0)$ $(\lambda_1=\lambda_2=1)$.[1] Hazards are zero over $(.5, t_0-.5)$.

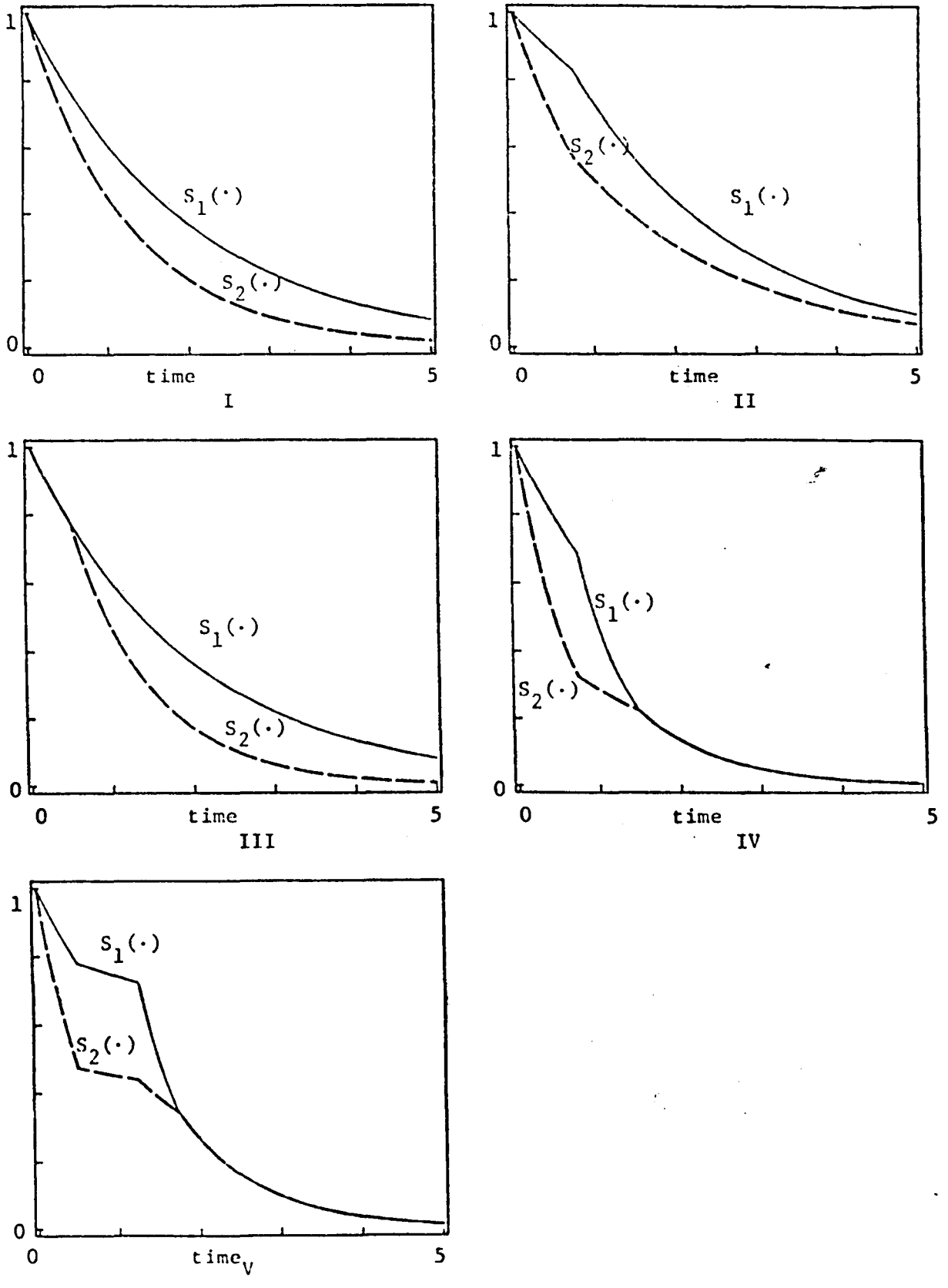


Figure 3

Piecewise exponential stochastic ordering survival configurations.

Table 1

Size simulation results from 5000 replications with equal censoring, $n_1 = n_2 = 20$.

Survival	Censoring	% Censored	$\alpha = .05$						$\alpha = .01$					
			SW_c^p	SW_c^{up}	SW_{Vc}^p	SW_{Vc}^{up}	Lgk	Wix	SW_c^p	SW_c^{up}	SW_{Vc}^p	SW_{Vc}^{up}	LgK	Wix
$S_{1,.5}$	U(0,1)	53	.047	.051	.047	.051	.051	.048	.010	.016	.010	.016	.011	.011
	U(0,2)	41	.049	.055	.048	.055	.050	.047	.009	.012	.009	.013	.010	.010
	U(0,3)	34	.055	.061	.055	.064	.055	.055	.010	.015	.010	.017	.012	.009
	Uncensored	0	.046	.053	.046	.053	.057	.052	.003	.004	.003	.004	.012	.011
$S_{1,1}$	U(0,1)	63	.049	.055	.050	.057	.050	.049	.010	.014	.010	.016	.011	.010
	U(0,2)	43	.053	.057	.053	.059	.051	.049	.009	.014	.009	.016	.011	.009
	U(0,3)	32	.055	.060	.056	.063	.055	.052	.011	.016	.010	.017	.013	.010
	Uncensored	0	.051	.054	.051	.054	.058	.053	.008	.012	.008	.012	.014	.011
$S_{1,2}$	U(0,1)	75	.053	.059	.055	.063	.053	.052	.009	.016	.010	.018	.009	.007
	U(0,2)	44	.046	.055	.048	.056	.044	.047	.008	.014	.008	.014	.012	.008
	U(0,3)	30	.054	.061	.054	.062	.057	.053	.009	.015	.010	.014	.013	.011
	Uncensored	0	.055	.060	.055	.060	.065	.056	.010	.014	.010	.014	.014	.011
$S_{1,3}$	U(0,1)	81	.053	.067	.055	.069	.053	.052	.008	.020	.010	.023	.011	.009
	U(0,2)	45	.052	.060	.053	.062	.053	.050	.010	.017	.010	.017	.011	.010
	U(0,3)	30	.054	.058	.055	.059	.056	.056	.011	.016	.011	.017	.015	.013
	Uncensored	0	.056	.060	.056	.060	.065	.056	.010	.013	.010	.013	.015	.012

Table 2

Power simulation results under Weibull proportional hazards.
1000 replications, equal censoring, $n_1 = n_2 = 20$

		α=.05						α=.01				
Censoring	Survival	SW^p_c	SW^p_{Vc}	Lgk	Wix	EE(SW_{Vc},Lgk)	Survival	SW^p_c	SW^p_{Vc}	Lgk	Wix	EE(SW_{Vc},Lgk)
U(0,1)	$S_1=S_{1,.5}$	.348	.352	.361	.330	96.4%	$S_1=S_{1,.5}$	.438	.436	.459	.382	95.0%
U(0,2)	$S_2=S_{\frac{1}{3},.5}$	.417	.418	.425	.393	92.5%	$S_2=S_{.167,.5}$	.474	.475	.511	.448	93.0%
U(0,3)	γ=1.73	.423	.430	.421	.389	102.1%	γ=2.45	.523	.520	.557	.476	93.4%
Uncensored		.429	-----	.527	.453	81.4%		.268	-----	.647	.532	41.4%
U(0,1)	$S_1=S_{1,1}$	.432	.440	.432	.424	101.9%	$S_1=S_{1,1}$	.308	.325	.352	.293	92.3%
U(0,2)	$S_2=S_{\frac{1}{2},1}$	.523	.529	.548	.491	96.5%	$S_2=S_{.4,1}$	.448	.450	.485	.399	92.8%
U(0,3)	γ=2.00	.573	.578	.574	.518	100.7%	γ=2.5	.534	.534	.565	.470	94.5%
Uncensored		.675	-----	.691	.588	97.7%		.532	-----	.654	.532	81.3%
U(0,1)	$S_1=S_{1,2}$	.371	.397	.409	.383	97.1%	$S_1=S_{1,2}$	.273	.293	.313	.290	93.6%
U(0,2)	$S_2=S_{\frac{1}{3},2}$	.599	.613	.630	.550	97.3%	$S_2=S_{.606,2}$	.452	.464	.514	.422	90.3%
U(0,3)	γ=2.25	.656	.668	.689	.598	97.0%	γ=2.72	.550	.551	.602	.516	91.5%
Uncensored		.770	-----	.803	.703	95.9%		.667	-----	.734	.608	90.9%
U(0,1)	$S_1=S_{1,3}$	.330	.362	.372	.348	97.3%	$S_1=S_{1,3}$	.173	.202	.223	.203	90.6%
U(0,2)	$S_2=S_{\frac{1}{3},3}$	.603	.614	.652	.575	94.2%	$S_2=S_{.714,3}$	.470	.481	.542	.447	88.7%
U(0,3)	γ=2.37	.663	.667	.719	.628	92.8%	γ=2.75	.500	.505	.568	.477	88.9%
Uncensored		.796	-----	.851	.765.	93.5%		.654	-----	.740	.620	88.4%

Table 3

Power simulation results for the configurations of Figure 2.

500 replications, $n_1=n_2=50$, $\alpha=.05$, and equal truncation censoring $C_i(t)=I\{t<t_0\}$, i=1,2.

t_0	SW_c^p	Lgk	Wix
1	.694	.510	.622
1.5	.748	.512	.612
2.5	.754	.516	.600
3.5	.746	.474	.590

Table 4

Power simulation results for configurations I through V of Figure 3. 500 replications, $n_1=n_2=50$ and equal censoring.

	U(0,2) Censoring				No Censoring		
Survival	SW_c^p	$SW_{\underline{V}c}^p$	Lgk	Wix	SW_c^p	Lgk	Wix
I	.472	.504	.520	.488	.836	.822	.726
II	.804	.772	.772	.788	.404	.400	.722
III	.210	.242	.274	.222	.824	.828	.580
IV	.828	.750	.688	.858	.420	.408	.858
V	.898	.868	.776	.866	.432	.242	.576