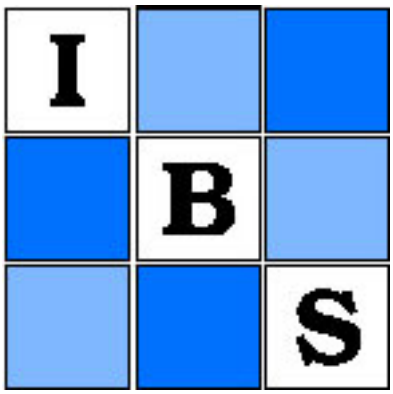

BIOMETRICS 45, 497–507
June 1989

Weighted Kaplan–Meier Statistics: A Class of Distance Tests for Censored Survival Data

Margaret Sullivan Pepe and Thomas R. Fleming

Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, Washington 98104, U.S.A.

and

University of Washington, Department of Biostatistics, SC-32, Seattle, Washington 98195, U.S.A.

SUMMARY

A class of statistics based on the integrated weighted difference in Kaplan–Meier estimators is introduced for the two-sample censored data problem. With positive weight functions these statistics are intuitive for and sensitive against the alternative of stochastic ordering. The standard weighted log-rank statistics are not always sensitive against this alternative, particularly if the hazard functions cross.

Qualitative comparisons are made between the weighted log-rank statistics and these weighted Kaplan–Meier (WKM) 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 favorably with the log-rank procedure even under the proportional hazards alternative, and may perform better than it under the crossing hazards alternative.

1. Introduction

Consider the classical two-sample censored data survival analysis problem, with survival continuous and censoring independent of survival in each group. The question to be addressed is whether survival in group 1 is better than survival in group 2. If the survival functions cross, then a summary measure of survival on which to base the comparison is needed. The mean survival time and average hazard rate are natural choices. If survival in the two groups is stochastically ordered, however,

$$S_1(t) \geqslant S_2(t) \quad \text{for all } t, \quad S_1(\cdot) \neq S_2(\cdot), \tag{1.1}$$

then clearly population 1 has the better survival. The comparison in this case does not need to rely on an arbitrary summary measure. Indeed, any test procedure used to compare two groups, regardless of which summary measure on which it is based, should at least be sensitive to the alternative of stochastic ordering.

The commonly used test statistics for censored data are the log-rank (Cox, 1972) and Wilcoxon (Peto and Peto, 1972) statistics. It can be shown that these statistics essentially estimate integrated weighted differences in hazard functions $\sqrt{n}\int_0^\infty K(u)[\lambda_2(u) - \lambda_1(u)]\,du$ with $K(\cdot)$ positive (Gill, 1980). Thus, these tests are sensitive to alternatives of ordered hazard functions, though not necessarily to the more general alternative (1.1). Work by Breslow, Edler, and Berger (1984) and more recently by Fleming, Harrington, and

Key words: Kaplan–Meier estimators; Random censorship; Stochastic ordering; Two-sample problem.

O'Sullivan (1987) has focused on versatile test procedures that are sensitive to both the ordered hazards and crossing hazards alternatives. These procedures, like the log-rank and Wilcoxon tests, are essentially based on estimates of the hazard functions. Here, we present a class of statistics which are based directly on the estimated survival functions and which have been motivated by the alternative of stochastic ordering.

2. Motivation

2.1 *Weighted Kaplan–Meier* (WKM) *Statistics*

The form of the stochastic ordering alternative suggests a natural (albeit naive) measure of the difference in survival between the two groups, namely $\mu = \int_0^{t_0} [S_1(t) - S_2(t)]\, dt$, where t_0 is the length of the study period. We can interpret μ/t_0 as the average difference in the survival probabilities over the study period. We can also interpret μ as the difference in mean survival times when distributions are truncated at t_0, although it is not this interpretation that motivates μ as a measure for comparison under the general stochastic ordering alternative. If $\hat{S}_i(\cdot)$, $i = 1, 2$, are the Kaplan–Meier estimators of the survival functions (Kaplan and Meier, 1958), then a natural statistic on which to base a test procedure might be

$$\hat{\mu} = \int_0^{t_0} [\hat{S}_1(t) - \hat{S}_2(t)]\, dt.$$

In censored data it is well known that the Kaplan–Meier estimator $\hat{S}_i(t)$ can be very unstable for t close to t_0 in the presence of heavy censoring. Indeed, the limiting variance of $\sqrt{n}[\hat{S}_i(t) - S_i(t)]$ is given by

$$v_i(t) = S_i^2(t) \int_0^t \frac{1}{S_i(u)[C_i^-(u)]}\, \lambda_i(u)\, du,$$

where $C_i^-(u)$ is the probability of not being censored before time u (Gill, 1980). In many applications censoring is determined largely by the timing of subject entry into the study, with the probability of early entry being quite small. That is, $C_i^-(t)$ is small for t close to t_0. In addition, it will most often be the case that there are positive probabilities of failing near t_0 and of surviving the length of the study, i.e., $\lambda_i(t) > 0$ for t near t_0 and $S_i(t_0) > 0$ so that $v_i(t)$ can be large for t close to t_0. Consequently, μ is unstable in many applications of practical interest, and hence unsuitable as a test statistic.

To remedy such instability we introduce a random weight function $\hat{w}(\cdot)$ estimating a deterministic function $w(\cdot)$, which downweights the contributions of the $\hat{S}_1(t) - \hat{S}_2(t)$ in $\hat{\mu}$ over later time periods if censoring is heavy. We will show that if the weight function is chosen appropriately the resulting statistic will be stable.

Formally, we define a weighted Kaplan–Meier or WKM statistic as

$$\text{WKM} = \sqrt{\frac{n_1 n_2}{n}} \int_0^{T_c} \hat{w}(t)[\hat{S}_1(t) - \hat{S}_2(t)]\, dt,$$

where $T_c = \sup\{t\colon \hat{C}_1(t) \wedge \hat{C}_2(t) > 0\}$, $\wedge$ denotes minimum, n_i is the sample size in group i, $n = n_1 + n_2$, and $\hat{C}_i(\cdot)$ is the Kaplan–Meier estimator of the censoring survival function in group i. Stability conditions, given in the next section, will require $\hat{w}(t) = 0$ if $\hat{C}_i(t) = 0$, $i = 1$ or 2, so that we can in fact replace T_c with t_0 (or ∞). We include this endpoint T_c in the definition, however, for the sake of comparison with the endpoint used by the weighted log-rank statistics.

2.2 *Comparison with Weighted Log-Rank Statistics*

We can write a weighted log-rank statistic (or L_K statistic) as

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

where $N_i(t)$ is the process that counts the observed deaths in group i through time t, $Y_i(t)$ is the number of subjects still at risk at t, $\hat{K}(t)$ is a weight function, and $T = T_1 \wedge T_2$, where T_i is the last observation (censored or not) in group i. Note that T is the natural endpoint here since L_K is based on the difference in hazard functions, which can be estimated only while subjects are at risk in both groups, i.e., on $[0, T]$. If T is a censored observation then T_c and T are the same, but if T is an uncensored observation, then $T < T_c$ and the WKM statistic uses information from more of the time axis than does the weighted log-rank.

We have motivated the class of WKM statistics from the point of view of stochastic ordering alternatives. WKM statistics are based directly on estimates of the survival functions, in contrast to L_K statistics, which are based on estimates of the hazard functions. Another essential difference is that, unlike L_K statistics, WKM statistics are not generalized rank statistics. Since an L_K statistic is based on ranks, its power against a specific alternative might not be sensitive to the magnitude of the difference in survival time. To illustrate this point we consider the two (admittedly extreme but illustrative) configurations depicted in Figure 1. Configuration (b) differs from (a) only in that survival in group 1 has been shifted to the right by the addition of a lag time. Although the magnitude of the difference in survival is larger in (b) than it is in (a), because the ranks of the observations in the combined samples are unchanged from realization (a) to realization (b), a rank statistic yields exactly the same P-value for both realizations. In contrast, the non–rank-based WKM statistics should be much more sensitive to configuration (b) than to (a). Indeed, a WKM statistic is essentially the difference of two generalized L-statistics

$$\text{WKM} = -\int_0^{T_c} [\hat{S}_1(t) - \hat{S}_2(t)]\, d\left[\int_t^{T_c} \hat{w}(u)\, du \right]$$

$$= \int_0^{T_c} \left[\int_t^{T_c} \hat{w}(u)\, du \right] d[\hat{S}_2(t) - \hat{S}_1(t)],$$

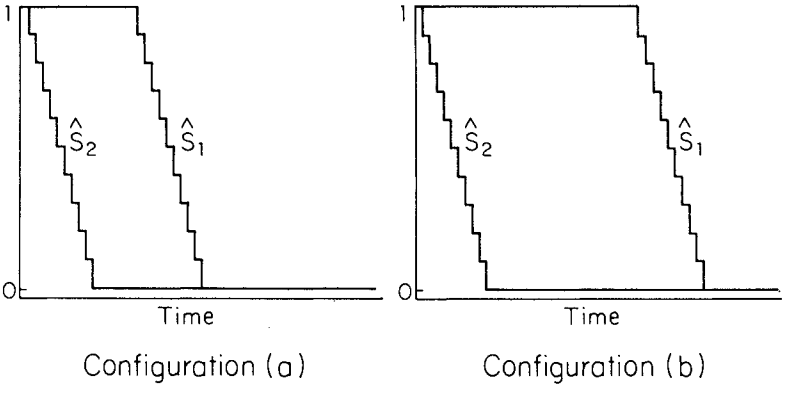

Figure 1. An example where rank statistics yield the same P-value for both a large and a small difference in survival time.

and hence the power of the test procedure is inherently dependent on the magnitude of the difference in survival time on some scale.

2.3 *Other Related Procedures*

Some nonparametric statistics based directly on the survivor functions have been proposed for this problem. Notably, generalizations of the Kolmogorov–Smirnov statistic to censored data have been studied by Fleming et al. (1980) and by Schumacher (1984). This statistic is based on the maximum distance between the two survivor functions and, though it may be very sensitive to a difference that is large but evident only over a short period of time, it can be very insensitive to a moderate difference that extends over a long period of time. In practice, the latter rather than the former will be of more clinical interest. Note also that because the Kolmogorov–Smirnov statistic is rank-based, it shares some of the shortcomings of L_K statistics.

A statistic closer to the WKM in philosophy is the Cramér–von Mises statistic, which in uncensored data can be written as

$$\text{CvM} = \int K(t)[\hat{S}_1(t) - \hat{S}_2(t)]^2 \, d\hat{F}(t),$$

where $\hat{F}(\cdot)$ is the pooled empirical distribution function and $K(\cdot)$ is a weight function. This is a weighted average of the squared distances between the estimated survival functions and is oriented toward the two-sided alternative $S_1(\cdot) \neq S_2(\cdot)$. A generalization to censored data was studied by Schumacher (1984). The Cramér–von Mises statistic differs from the WKM statistic in that again it is a rank statistic, the integration with respect to $\hat{F}(\cdot)$ allowing mass only at observed death times. Interestingly, the original statistic proposed by Cramér was $\int [\hat{S}_1(t) - \hat{S}_2(t)]^2 \, dt$, a two-sided version of an (unweighted) WKM statistic. Von Mises introduced the integration with respect to $\hat{F}$ to yield a distribution-free rank statistic, allowing for exact nonparametric testing in small samples. The WKM statistics proposed here retain the flavor of Cramér's original statistic. Although they are not distribution-free, we will choose the weight function so that the statistics are nonparametric in the sense that asymptotically valid tests can be performed without assumptions regarding the underlying survival and censoring distributions.

3. An Asymptotically Valid Test Procedure

Suppose the null hypothesis H_0: $S_1(\cdot) = S_2(\cdot) = S(\cdot)$ holds. Also assume that $\lim_{n\to\infty} n_i/n = p_i > 0$, i.e., both groups are a nonnegligible fraction of the total sample, and that the observation times are bounded. Asymptotically the relevant time interval for comparison is $[0, \tau]$ where $\tau = \sup\{t: S(t) \wedge C_1(t) \wedge C_2(t) \geq 0\}$. We will consider only random weight functions $\hat{w}(\cdot)$ that are good estimators of some $w(\cdot)$ in the sense that

$$\sup_{(0,\tau)} |\hat{w}(t) - w(t)| \xrightarrow{\text{P}} 0. \tag{3.1}$$

Natural weight functions which we have considered involve Kaplan–Meier estimators of the underlying survival and censoring distribution functions. Hence, consistency of the Kaplan–Meier estimator (Shorack and Wellner, 1986) will often be useful in verifying (3.1).

The real constraint on the weight function, to ensure stability of the WKM statistic, is that for some positive constants Γ and δ,

$$|w(t)| \leq \Gamma[C_i^-(t)]^{(1/2)+\delta} \quad \text{and} \quad |\hat{w}(t)| \leq \Gamma[\hat{C}_i^-(t)]^{(1/2)+\delta} \tag{3.2}$$

for all $t < \tau$ and $i = 1, 2$. Thus, we require that the weight function be small toward the end of the observation period if censoring is heavy, that is, if $\sqrt{n}[\hat{S}_1(t) - \hat{S}_2(t)]$ has a large

variance. In unpublished work, O'Sullivan and Fleming (Technical Report No. 163, Center for Stochastic Processes, University of North Carolina, 1986) show that if (3.2) is satisfied, then

$$\text{WKM} = \sqrt{\frac{n_1 n_2}{n}} \int_0^{T_c} \hat{w}(t)[\hat{S}_1(t) - \hat{S}_2(t)]\, dt \xrightarrow{\text{d}} \text{N}(0, \sigma^2), \tag{3.3}$$

where

$$\sigma^2 = -\int_0^{\tau} \frac{[\int_t^{\tau} w(u)S(u)\, du]^2}{S^2(t)} \left(\frac{p_1 C_1^{-1}(t) + p_2 C_2^{-}(t)}{C_1^{-}(t) C_2^{-}(t)} \right) dS(t).$$

The variance expression σ^2 is not intuitive, though in the simple uncensored data case with $w(\cdot) = 1$, it can be verified that σ^2 is the variance of the distribution function $1 - S(\cdot)$ (O'Sullivan, unpublished Ph.D. thesis, University of Washington, 1986). We would expect this since we can rewrite WKM in this case as $\sqrt{n_1 n_2/n}(\bar{X}_1 - \bar{X}_2)$, where $\bar{X}_i$ is the sample mean. Two natural estimators of σ^2, an unpooled $\hat{\sigma}^2_{up}$ and a pooled $\hat{\sigma}^2_p$, are

$$\hat{\sigma}^2_{up} \equiv -\sum_{i=1}^{2} \hat{p}_{3-i} \frac{n_i}{n_i - 1} \int_0^{T_c} \frac{[\int_t^{T_c} \hat{w}(u)\hat{S}_i(u)\, du]^2}{\hat{S}_i(t)\hat{C}_i^{-}(t)\hat{S}_i^{-}(t)} d\hat{S}_i(t)$$

and

$$\hat{\sigma}^2_p \equiv -\int_0^{T_c} \frac{[\int_t^{T_c} \hat{w}(u)\hat{S}(u)\, du]^2}{\hat{S}(t)\hat{S}^{-}(t)} \frac{\hat{p}_1 \hat{C}_1^{-}(t) + \hat{p}_2 \hat{C}_2^{-}(t)}{\hat{C}_1^{-}(t)\hat{C}_2^{-}(t)} d\hat{S}(t),$$

where $\hat{S}(\cdot)$ is the Kaplan–Meier estimator calculated from the pooled sample and $\hat{p}_i = n_i/n$. These are consistent under the null hypothesis (O'Sullivan, unpublished Ph.D. thesis, 1986). The unpooled estimator is also consistent under a fixed alternative, estimating the two components of the variance separately. In the classical uncensored case with $\hat{w}(\cdot) = 1$, $\hat{\sigma}^2_{up} = (n_2/n)S_1^2 + (n_1/n)S_2^2$, where S_i^2 is the sample variance from the ith sample, so that $\hat{\sigma}^2_{up}$ is the usual unpooled variance estimator. On the other hand, $\hat{\sigma}^2_p$ is the sample variance from the two samples combined, and is not the usual pooled variance estimator used with the difference in sample means.

4. The Choice of Weight Function

The stability constraint (3.2) requires that $\hat{w}$ be a function of $\hat{C}_1^{-}$ and $\hat{C}_2^{-}$. For example,

$$\hat{w}_c(t) = \frac{\hat{C}_1^{-}(t)\hat{C}_2^{-}(t)}{\hat{p}_1 \hat{C}_1^{-}(t) + \hat{p}_2 \hat{C}_2^{-}(t)},$$

which is akin to a geometric average of the two censoring survivor function estimators, and $[\hat{w}_c(\cdot)]^{(1/2)+\delta}$, $\delta > 0$, satisfy (3.2). Since these weight functions reduce to unity in uncensored data, the resulting WKM test statistics can be regarded as generalizations of the z-test statistic to censored data. It is desirable that the weight function render the statistic interpretable as well as stable so that the power of the test procedure depends on some population parameter of interest. If the difference in mean survival time is a parameter of interest, for example, then $\hat{w}_c(\cdot)$ is a natural choice. In uncensored data this WKM estimates the mean difference in survival during the study period, and under stochastic ordering in censored data a lower bound is estimated.

In a medical context the weight function $w(\cdot)$ might involve some notion of quality of life. For example, in the comparison of the effects on survival of two aggressive treatments administered over a short time interval $[0, t_0]$, only the long-term quality lifetime gained

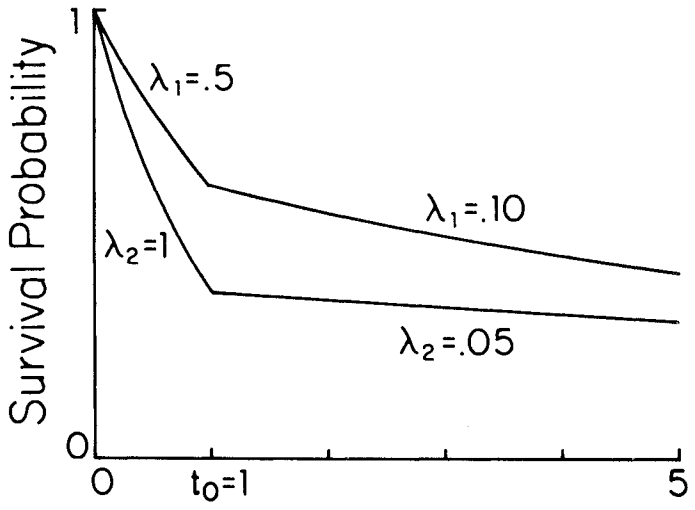

Figure 2. An example where group 1 has better long-term survival probabilities but higher long-term hazard rates than group 2.

during ($t \geq t_0$) for one treatment over the other may be relevant. For the choice $w(t) = 0$, $t \leq t_0$, and $w(t) = 1$, $t > t_0$, $\sqrt{n/(n_1 n_2)}$ WKM is an estimate of the mean difference in quality lifetime over the study period. In censored data one might use $\hat{w}_c(t)I\{t \geq t_0\}$, which provides a lower bound on this difference under stochastic ordering. This very simple situation cannot be accommodated by statistics based on the hazard function. For example, in Figure 2, group 1 is seen to have more quality survival time, yet the hazard in group 1 is larger than the hazard in group 2 during the relevant time period $t \geq t_0$.

In marketing settings the notion of cost will enter naturally into the weight function. For example, in a wood-processing plant consider the comparison of two chemical treatments to enhance the strength of wood planks. A gradually increasing load applied to a plank yields a "load at breakage" (survival time) random variable. If $V(u)$ is the probability that a plank will be used at a load less than u in practice, then the expected proportion of breakages in practice is

$$\int_0^\infty [1 - S(u)]\, dV(u) = \int_0^\infty v(u)[1 - S(u)]\, du.$$

In terms of real cost the natural statistic on which to base a comparison is

$$\int_0^\infty \hat{v}(u)[\hat{S}_1(u) - \hat{S}_2(u)]\, du.$$

In censored data the weight function $\hat{w}_c(u)\hat{v}(u)$ might be used.

In these two examples, the quantities to be estimated are essentially differences of two weighted (survival time) averages, or location estimates. These are well estimated in uncensored data. Stable estimation in censored data, however, requires downweighting of information over periods of heavy censoring. Hence, censored data are intrinsically unsuitable for estimation of these "real-time" parameters. However, realizing the limitations of the data, we can develop test procedures which, by virtue of the particular choice of weight function, are sensitive to the alternative of most interest. These types of techniques have not been considered in survival analysis before.

5. Small-Sample Simulation Results

In the simulation studies we considered the weight functions $\hat{w}_c$ and $\sqrt{\hat{w}_c}$ because they will often be natural choices in practice. A subscript c or $\sqrt{c}$ denotes the corresponding WKM statistic. A superscript p (pooled) or up (unpooled) indicates the variance estimator used. The simulations were performed on an IBM-AT personal computer, using the APL programming language and its inherent random number generator. All tests were one-sided.

5.1 *Size Properties*

The survival distributions chosen were Weibull, $S_{b,a}(t) = e^{-(t/b)^a}$, $t \geq 0$, with $b = 1$ and $a =$ .5, 1, 2, and 3. This group of survival functions is diverse in terms of skewness and tail-weight, factors that might be expected to affect the empirical sizes of the tests under study. The censoring distributions were equal and uniform U(0, c). 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 WKM_c and $\mathrm{WKM}_{\sqrt{c}}$ provide equally acceptable empirical significance levels, with the test based on $\mathrm{WKM}_{\sqrt{c}}$ being slightly more anticonservative than WKM_c in general. From a practical point of view, however, we can conclude that $\mathrm{WKM}_{\sqrt{c}}$ is equally as acceptable to use in small samples as WKM_c. Indeed, the simulation results suggest that the asymptotic result (3.3) may also be valid for $\mathrm{WKM}_{\sqrt{c}}$ despite the fact that (3.2) does not hold.

Table 1 also indicates the superior performance of the pooled variance estimator over the unpooled estimator, tests based on $\hat{\sigma}^2_{up}$ being unacceptably anticonservative, especially at lower P-values. Further simulations (presented in O'Sullivan, unpublished Ph.D. thesis, University of North Carolina, 1986) indicate that the situation worsens considerably for $\hat{\sigma}^2_{up}$ in unequally censored data and yet $\hat{\sigma}^2_p$ performs very well. Reasons for this have yet to be explored. We recommend use of $\hat{\sigma}^2_p$ in practice and the remaining results presented here pertain only to procedures using this variance estimator.

The empirical levels of WKM^p_c and $\mathrm{WKM}^p_{\sqrt{c}}$ are both very close to the nominal level across a broad range of situations. Only with uncensored heavy-tailed data do the sizes deviate substantially from the nominal level and then only at the lower .01 level rather than at the .05 level. Since in uncensored data we are essentially dealing with the z-test and intuitively an outlier will influence the sample variance considerably more than it will affect the sample mean, it is not surprising that the tests are conservative in this case. Benjamini (1983) has shown that the conservatism occurs primarily at low P-values. To protect against such conservatism, classical statistical methodology would suggest "trimming the data," using trimmed sample means for comparison rather than true sample means. That is to say, artificial censoring of the data to increase robustness of the test procedure is suggested! The natural censoring seen in survival data should yield WKM^p statistics robust in size in real applications.

5.2 *Power*

The results of some simulations under the various stochastic ordering configurations of Figure 3 are given in Table 2. Configurations I(a), (b), (c), and (d) are Weibull proportional hazards alternatives, with $a = .5$, 1, 2, and 3, respectively. The scale parameter b for S_1 is 1 in all cases. In censored data $\mathrm{WKM}_{\sqrt{c}}$ is slightly more powerful than WKM_c, primarily because the largest difference in the survival functions occurs toward the end of the censoring distribution and $\mathrm{WKM}_{\sqrt{c}}$ puts relatively more weight there than does WKM_c.

Table 1

Size simulation results from 5,000 replications with equal censoring, $n_1 = n_2 = 20$

			$\alpha = .05$						$\alpha = .01$					
Survival	Censoring	% censored	WKM_c^p	WKM_c^{up}	$WKM_{\sqrt{c}}^p$	$WKM_{\sqrt{c}}^{up}$	Log-rank	Wilc.	WKM_c^p	WKM_c^{up}	$WKM_{\sqrt{c}}^p$	$WKM_{\sqrt{c}}^{up}$	Log-rank	Wilc.
	U(0, 1)	53	.047	.051	.047	.051	.051	.048	.010	.016	.010	.016	.011	.011
	U(0, 2)	41	.049	.055	.048	.055	.052	.047	.009	.012	.009	.013	.010	.010
	U(0, 3)	34	.055	.061	.055	.064	.055	.055	.010	.015	.010	.017	.012	.009
$S_{1,.5}$	Uncensored	0	.046	.053	.046	.053	.057	.052	.003	.004	.003	.004	.012	.011
	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
$S_{1,1}$	Uncensored	0	.051	.054	.051	.054	.058	.053	.008	.012	.008	.012	.014	.011
	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
$S_{1,2}$	Uncensored	0	.055	.060	.055	.060	.065	.056	.010	.014	.010	.014	.014	.011
	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
$S_{1,3}$	Uncensored	0	.056	.060	.056	.060	.065	.056	.010	.013	.010	.013	.015	.012

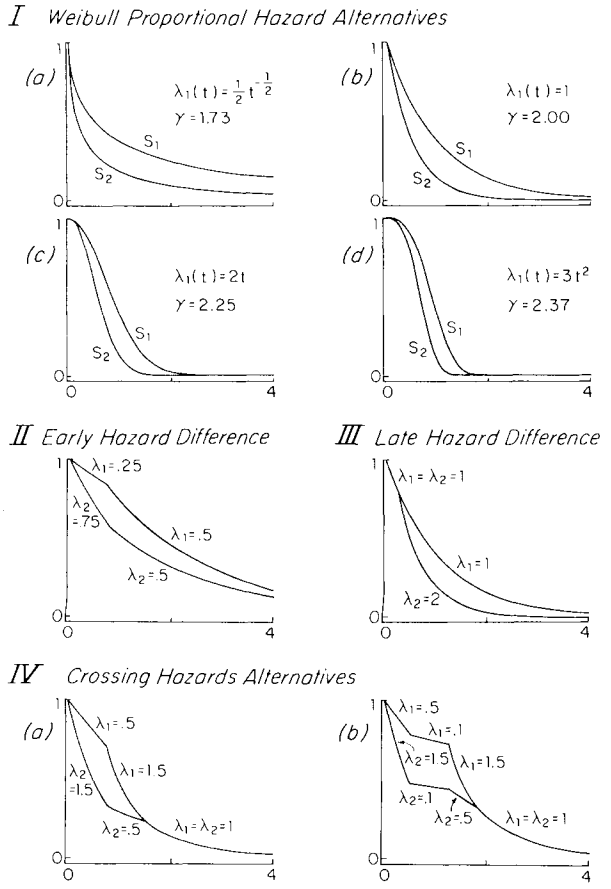

Figure 3. Stochastic ordering alternatives considered in the power simulations. Configurations in I are Weibull with constant hazard ratio γ. Configurations in II, III, and IV are piecewise exponential.

The difference, however, is very small. The log-rank is known to be the locally most powerful test statistic against proportional hazards alternatives when the censoring distributions are equal (Gill, 1980) and indeed its efficiency in this situation is a major reason for its popularity. Table 2 indicates that across a broad range of proportional hazards alternatives the WKM_c and $WKM_{\sqrt{c}}$ statistics attain high efficiency. Only in the uncensored heavy-tailed configuration I(a) does the log-rank gain substantially over the WKM statistics. This most likely can be explained by the conservatism of this test in small samples under the null hypothesis in uncensored heavy-tailed data. In censored data it appears that WKM_c and $WKM_{\sqrt{c}}$ compare well with the most efficient log-rank statistic under the proportional hazards alternatives considered here.

If groups 1 and 2 are regarded as treatment and control groups, respectively, then the treatment decreases the constant hazard rate uniformly over time in the cases studied above. Often a treatment will decrease the hazard for some initial period but its effect on the hazard becomes negligible later on, as in configuration II. Since the power of a weighted log-rank statistic is governed by $\int_0^{\tau} K(t)[\lambda_1(t) - \lambda_2(t)]\, dt$, the later time period contributes nothing to the power of these statistics, even though differences in survival functions remain and presumably add to the power of WKM. Indeed, for the particular configuration chosen, the WKM statistics are seen to perform better than the log-rank, though not substantially

Table 2
Power simulation results at significance level $\alpha = .05$ *for configurations of Figure* 3, 500 *replications*

	Uniform (0, 2) censoring				No censoring		
Survival	WKM_c	$WKM_{\sqrt{c}}$	Log-rank	Wilc.	WKM_c	Log-rank	Wilc.
I[a] (a)	.417	.418	.425	.393	.429	.527	.453
(b)	.523	.529	.548	.491	.675	.691	.588
(c)	.599	.613	.630	.550	.770	.803	.703
(d)	.603	.614	.652	.575	.796	.851	.765
II[b]	.804	.772	.722	.788	.404	.400	.722
III[b]	.520	.548	.548	.378	.850	.844	.556
IV[b] (a)	.828	.750	.688	.858	.420	.408	.858
(b)	.898	.868	.776	.866	.432	.242	.576

[a] $n_1 = n_2 = 20$.
[b] $n_1 = n_2 = 50$.

better since later differences in survival are small. The generalized Wilcoxon statistic, which relative to the log-rank places more weight at early rather than at late hazard differences, is quite well suited to the alternative of configuration II.

In configuration III the effect of the treatment on the hazard does not manifest itself until later in time. The Wilcoxon performs poorly though both the log-rank and WKM statistics perform very well. Alternatives where long-term differences in the survival functions occur are often of particular interest. The low power of the Wilcoxon in such situations is a reason for its recent loss in popularity.

The hazard functions cross in configuration IV, the treatment decreasing the hazard early in time but increasing the hazard late in time. The log-rank performs poorly and the WKM can perform much better than it in such situations. The Wilcoxon can also perform well in such situations because it places little weight on the late negative differences $\lambda_1(t) - \lambda_2(t)$. However, an intermediate period of approximately equal hazards, which may well occur in practice, can detract from its efficiency as in IV(b).

We can draw the following general conclusions from the simulation studies. First, the power of WKM_c and $WKM_{\sqrt{c}}$ is determined by the magnitude of the difference in the observed survival time, scaled by the overall variability. In contrast, the power of a weighted log-rank statistic is governed by the difference in the hazard functions. Second, across a broad range of stochastic ordering alternatives the WKM statistics are good competitors to the log-rank, even under the proportional hazards alternative. They can perform substantially better than the log-rank when the hazard functions cross.

6. Discussion

A basic assumption made for the development of asymptotic distribution theory is that $S(\cdot)$ is continuous. Although survival time may in truth have a continuous distribution, in practice data are always recorded in discrete units. Hence, ties will often occur in real data. O'Sullivan (unpublished Ph.D. thesis, University of North Carolina, 1986) has shown that if the recording unit is small and censored observations occur after survival time observations at tied data points, then the statistic calculated from its definition (with the ties incorporated in the usual definition of the Kaplan–Meier estimators, etc.) is a close approximation to that calculation had the true observation times been recorded. Since WKM statistics are real-time statistics, this is to be expected.

In summary, we have introduced a class of nonparametric statistics that seem intuitive for the general alternative of stochastic ordering. In contrast to the classical nonparametric

statistics for the censored data problem, which are generalized rank statistics, WKM statistics are generalizations of location test statistics. In some cases the weight function can be chosen so that the test statistic is an estimator of some population parameter of interest. With weight function $\hat{w}_c$ the statistic is a generalization of the z-test statistic to censored data. This test seems to compare favorably with the popular log-rank test statistic across a broad range of stochastic ordering alternatives.

Further work on the choice of weight functions that are optimal with respect to efficiency against particular families of alternatives is being done. Stratified test procedures, K-sample test procedures, and procedures based on the joint use of a weighted log-rank and a WKM statistic are also being investigated. Finally, some practical applications of WKM statistics will further illustrate their real worth, especially relative to the current popular log-rank and Wilcoxon procedures.

Acknowledgements

Research was supported in part by National Cancer Institute Grant CA32693. The authors wish to thank the reviewers for their helpful suggestions and Gary Longton for assistance in preparing the manuscript.

Résumé

Une classe de statistiques basées sur la différence intégrée pondérée dans les estimateurs de Kaplan–Meier est présentée pour le problème des données censurées sur deux échantillons. Avec des fonctions de poids positives ces statistiques sont intuitives et sensibles contre l'alternative d'un ordre stochastique. Les statistiques du lograng standard pondéré ne sont pas toujours sensibles contre cette alternative particulièrement si les fonctions de risque croissent.

Des comparaisons qualitatives sont faites entre les statistiques du lograng pondéré et ces statistiques de Kaplan–Meier pondérées. Une distribution théorique asymptotique nulle est spécifiée et le choix d'une fonction de pondération est discuté en détail. Des résultats d'études de simulation sur de petits échantillons indiquent que ces statistiques se comportent favorablement par rapport à la procédure du lograng meme sous l'alternative de risques proportionnels et peuvent etre supérieures sous une alternative de risques croissants.

References

Benjamini, Y. (1983). Is the t test really conservative when the parent distribution is long-tailed? *Journal of the American Statistical Association* **78,** 645–654.

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

Cox, D. R. (1972). Regression models and lifetables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–200.

Fleming, T. R., Harrington, D. P., and O'Sullivan, M. (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. *Journal of the American Statistical Association* **82,** 312–320.

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

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

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53,** 457–481.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank-invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A* **135,** 185–207.

Schumacher, M. (1984). Two-sample tests of the Cramér–von Mises and Kolmogorov–Smirnov type for randomly censored data. *International Statistical Review* **52,** 263–281.

Shorack, G. A. and Wellner, J. A. (1986). *Empirical Processes with Applications in Statistics.* New York: Wiley.

Received December 1986; *revised February and July* 1988.