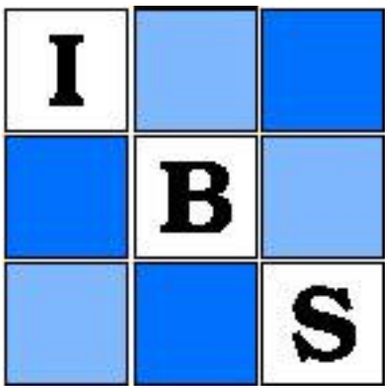

BIOMETRICS 45, 157–170
March 1989

A General Class of Nonparametric Tests for Survival Analysis

Michael P. Jones

Department of Preventive Medicine, College of Medicine,
University of Iowa, Iowa City, Iowa 52242, U.S.A.

and

John Crowley

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

SUMMARY

Tarone and Ware (1977, *Biometrika* **64,** 156–160) developed a general class of s-sample test statistics for right-censored survival data that includes the log-rank and modified Wilcoxon procedures. Subsequently, many authors have considered two- and s-sample classes in detail. In this paper a family of nonparametric statistics is shown to unify existing and generate new test statistics for the s (≥ 2)-sample, s-sample trend, and single continuous covariate problems.

1. Introduction

Many of the popular nonparametric two-sample test statistics for censored survival data, such as the log-rank (Mantel, 1966), generalized Wilcoxon (Gehan, 1965), and Peto–Peto (1972) test statistics, have been shown to be special cases of a general two-sample statistic, differing only in the choice of weight function (Tarone and Ware, 1977; Gill, 1980). This work has been extended to a general s-sample statistic (Tarone and Ware, 1977; Andersen et al., 1982) which includes the s-sample log-rank, Breslow (1970), generalized Wilcoxon, and Prentice (1978) linear rank statistics. There are, however, many survival analysis procedures that do not obviously fit into the general class above. Examples are statistics that deal with the s-sample trend problem, such as the Jonckheere (Gehan, 1965) and Tarone (1975) trend statistics, and statistics that deal with continuous, fixed, and time-dependent covariate problems, such as the Cox (1972) score test, the linear rank statistics of Prentice (1978), the modified Kendall rank correlation of Brown, Hollander, and Korwar (1974), and the logit rank test of O'Brien (1978). In this paper we propose a nonparametric statistic that is applicable to the s-sample, s-sample trend, and single continuous covariate problems, and as a general class unifies all of the aforementioned test procedures and some new statistics that are robust to outliers in the covariable space. In particular, the Kendall and Jonckheere statistics, whose textbook definitions are in terms of U-statistics, will be written as rank statistics. One of the new statistics is a generalization of the log-rank test appropriate for the single continuous covariate setting and robust to extreme covariates.

In a common approach to developing test procedures for survival data in the spirit of Mantel (1966), a contingency table is formed at each failure time containing the risk set, covariable information, and identification of the failing individual(s). Most of the ad hoc

Key words: Censored data; Kendall rank correlation; Log-rank test; Time-dependent covariates.

nonparametric procedures as well as Cox regression for time-dependent covariates (Cox, 1972, 1975) can be derived from these contingency tables. In Section 2 we present the contingency table approach to the survival problem, and from these tables we motivate a class of single-covariate nonparametric test statistics. In Section 3 we show that many of the common survival analysis tests are special cases of this general statistic. This class is also extended to the s-sample problem. In Section 4 we discuss several topics including stratification and small-sample properties, before summarizing the benefits of the proposed general class of procedures.

2. The Class of Statistics

Suppose there are k distinct failure times, $t_{(1)} < \cdots < t_{(k)}$, and that $t_{(0)} = 0$. Let $R(t_{(i)})$ denote the risk set at time $t_{(i)}$. Each of the $n(t_{(i)})$ individuals at risk at $t_{(i)}$ has a covariate $X_j(t_{(i)}), j \in R(t_{(i)})$. At time $t_{(i)}$, $b(t_{(i)})$ individuals fail. Each individual will eventually fail or be censored. Conditional on the covariate, the censoring and failure mechanisms are assumed to be mutually independent. For the rest of this paper, we shall assume all n individuals under study are at risk at time 0, i.e., $n = n(0)$.

Each of the $n(t)$ individuals at risk at time t has a quantitative label $Z_j(t), j \in R(t)$. We shall assume that the labels $\{Z_j(t), j \in R(t)\}$ have been constructed so that increasing labels correspond to either monotone increasing or monotone decreasing chances of failing, but consistently one or the other. This is our nonparametric model relating the label to the hazard of failure, typically called an ordered hazards model. The $Z_j(t)$ may represent the covariate values but can be more general and represent some function of the covariates, e.g., the ranks of the covariates. Define an at-risk indicator $Y_j(t) = 1$ if $j \in R(t)$ and 0 otherwise, and an observed failure time indicator $J_j(t) = 1$ if the jth person is observed to fail at time t and 0 otherwise. At time t the data can be summarized in a contingency table (Table 1).

Table 1
Summary of data at time t

Individual	Label	Indicator of failure	At risk
1	$Z_1(t)$	$J_1(t)$	$Y_1(t)$
2	$Z_2(t)$	$J_2(t)$	$Y_2(t)$
⋮	⋮	⋮	⋮
n	$Z_n(t)$	$J_n(t)$	$Y_n(t)$
Sum		$b(t)$	$n(t)$

The general class of statistics we propose is given by

$$T(w, Z) = \sum_{i=1}^{k} w(t_{(i)}) \sum_{j=1}^{n} J_j(t_{(i)})[Z_j(t_{(i)}) - \bar{Z}(t_{(i)})], \tag{1}$$

where $\bar{Z}(t) = n^{-1}(t) \sum Y_j(t)Z_j(t)$ is the average label at risk at time t, and $w(t)$ is a weight function. $T(w, Z)$ will often be abbreviated to T when the meaning is clear from the context. If the failure mechanism prefers large values of the labels, then T will be large; if it prefers small labels, T will be small. If we fix the labels and margins of Table 1, $(J_1(t), \ldots, J_n(t))$ has a multiple hypergeometric distribution under H_0 with mean $b(t)n^{-1}(t) \times (Y_1(t), \ldots, Y_n(t))$ and covariance matrix with (j, l)th element

$$b(t)[n(t) - b(t)]n^{-2}(t)[n(t) - 1]^{-1} Y_j(t)[\delta_{jl}n(t) - Y_l(t)],$$

where $\delta_{jl} = 1$ if $j = l$ and 0 otherwise. In this framework it is easy to verify that

$$\mathrm{E}\left[\sum_{j=1}^{n} J_j(t)Z_j(t)\right] = b(t)\bar{Z}(t), \tag{2}$$

$$\mathrm{var}\left[\sum_{j=1}^{n} J_j(t)Z_j(t)\right] = b(t)\,\frac{n(t) - b(t)}{n(t) - 1}\,C_{ZZ}(t), \tag{3}$$

where $C_{ZZ}(t) = n^{-1}(t) \sum Y_j(t)\,[Z_j(t) - \bar{Z}(t)]^2$ is the sample variance of the labels of those at risk at time t, in which $n^{-1}(t)$ is used rather than $[n(t) - 1]^{-1}$. Let $b^*(t) = b(t)[n(t) - b(t)]/[n(t) - 1]$. When there are no tied failure times, $b^*(t) \equiv 1$. As a variance estimator, therefore, let us propose

$$V(w, Z) = \sum_{i=1}^{k} w^2(t_{(i)})b^*(t_{(i)})C_{ZZ}(t_{(i)}). \tag{4}$$

One would hope to be able to refer the statistic $T/\sqrt{V}$ to a standard normal table to obtain a significance level for the association between the label and the hazard. Note that in (2) and (3) the mean and variance are functions of random variables. Furthermore, in the calculation of V, the hypergeometric variances are summed across tables as if they were independent. Of course, they are highly dependent. However, it still seems intuitively reasonable that one should be able to condition across time on previous events and proceed as above. In fact, many authors have used martingale theory to derive the statistical properties of tests whose motivation is based on changes in the risk set over time. Gill (1980) applied this theory to a class of two-sample test statistics with quite general weight function, and Andersen and Gill (1982) investigated the Cox relative risk regression. Using similar techniques, Jones and Crowley (Technical Report No. 88-1, Department of Preventive Medicine, University of Iowa, 1988) showed that under the null hypothesis and suitable regularity conditions, V is consistent for var(T) and that $T/\sqrt{V}$ is asymptotically a standard normal random variable for a general class of weight functions w and label functions Z. The failure time distribution need not be continuous.

The T statistic defined in (1) represents a broad class of nonparametric procedures whose individual members are specified when particular choices for the weight function w and label Z are made. We shall introduce seven different choices for the label $Z_j(t)$:

(L1) The most obvious choice of label is $Z_j(t) = X_j(t)$, in which case T is the Cox score statistic of H_0: $\beta = 0$. Since Cox regression is quite sensitive to outliers in the covariable space, appropriate alternative choices might be labels robust to such outliers.

(L2) Let $r_j(t)$ be the rank of $X_j(t)$ among $\{X_i(t)\colon i \in R(t)\}$; then $Z_j(t) = r_j(t)/n(t)$ would be a robust choice. Because of the difficulty in ranking a multidimensional point, this procedure is typically confined to a single covariate.

(L3) If $Z_j(t)$ is a function of $r_j(t)/n(t)$, such as normal scores or approximate normal scores, then one not only gains in robustness but also in approximate normality of T in small samples, where the central limit theorem does not hold. This should lead to an improvement in the Type I error rate in small studies.

(L4) If the covariate is truncated so that

$$Z_j(t) = \begin{cases} X_{\mathrm{L}} & \text{if } X_j(t) < X_{\mathrm{L}} \\ X_j(t) & \text{if } X_{\mathrm{L}} \leqslant X_j(t) \leqslant X_{\mathrm{U}}, \\ X_{\mathrm{U}} & \text{if } X_j(t) > X_{\mathrm{U}} \end{cases}$$

where X_{L} and X_{U} are predefined constants, then outliers are assigned values X_{L} and X_{U}, and the actual covariate data between X_{L} and X_{U} are used.

(L5) If the covariate values at risk $\{X_j(t): j \in R(t)\}$ are replaced by a Winsorized sample $\{Z_j(t): j \in R(t)\}$ in which α_U% of the upper tail have been moved down to the next observed covariate value, and similarly for α_L% of the lower tail, then once again most of the actual covariates are used with some protection against outliers.

(L6) A grouped covariate, e.g., $Z_j(t) = d_i$ when $x_{i-1} < X_j(t) \leq x_i$ where $i = 1, \ldots, s$ and $-\infty = x_0 < x_1 < \cdots < x_{s-1} < x_s = \infty$ are predefined constants, would provide protection against data that are only approximately correct. The constants $\{d_i\}$ might be chosen to represent any type of trend, e.g., linear ($d_i = i$), quadratic ($d_i = i^2$), or logarithmic ($d_i = \log i$), $i = 1, \ldots, s$. Alternatively, they could be midpoints of the intervals $[x_{i-1}, x_i]$ for $i = 2, \ldots, s-1$ with d_1 = some minimum and d_s = some maximum. The value of d_i may vary with time, e.g., the average or median covariate within $[x_{i-1}, x_i]$ among those at risk at time t.

(L7) Finally, consider $Z_j(t) = \Psi[X_j(t) - \bar{X}(t)]$, where Ψ downweights outliers similar to the robust Ψ function used in M-estimation procedures. The general form

$$\Psi(x) = \begin{cases} c + g(|x - c|) & \text{if } x > c \\ x & \text{if } |x| \leq c \\ -c - g(|x + c|) & \text{if } x < -c \end{cases}$$

for some nonnegative function g and constant c covers a wide class. For example, the function $g(|a|) = \log(1 + |a|)$ brings in outliers without bounding their contribution to T, whereas the function $g \equiv 0$ bounds the outlier contribution. Similarly, one could modify the truncated covariate with such a g function.

The weight function $w(t)$ is chosen to be sensitive to a prescribed alternative hypothesis. As seen in (1) when $w \equiv 1$, T weights failures evenly over time. For convenience, let $\hat{w}(t; \alpha, \beta) = [\hat{S}_{KM}(t)]^{\alpha}[1 - \hat{S}_{KM}(t)]^{\beta}$, where $\hat{S}_{KM}$ is the Kaplan–Meier estimator. Note that $\hat{w}(t; 1, 0)$ gives more weight to early failure times, $\hat{w}(t; 0, 1)$ to late failure times, and $\hat{w}(t; 1, 1)$ to mid failure times. Several popular forms of T employ $w(t) = n(t)$, the number at risk.

3. Equivalence to Popular Tests

In this section we show that various forms of the trend statistic T, specified by particular choices of the label and weight functions, are equivalent to many of the commonly used survival analysis procedures. The types of labels to be considered here are the raw covariate, the grouped covariate, the covariate rank, and the transformed rank of the covariate.

3.1 *Raw Covariate*

Cox relative risk regression based on the relative risk model $\lambda_j(t) = \lambda_0(t)\exp[\beta X_j(t)]$ (Cox, 1972) is the principal survival regression method based on covariates in use today. The Cox score statistic of H_0: $\beta = 0$ based on Breslow's (1974) approximate partial likelihood is

$$\sum_{i=1}^{k} [S(t_{(i)}) - b(t_{(i)})\bar{X}(t_{(i)})], \tag{5}$$

where $S(t)$ is the sum of the covariates of the failing individuals and $\bar{X}(t)$ is the average covariate at risk. The Cox score statistic (5) and T given by (1) are equivalent when $w(t) = 1$ and $Z = X$. Thus, the Cox score statistic can be written as $T(1, X)$.

Prentice (1978) developed another regression approach from the marginal likelihood of a rank vector of residuals using the accelerated failure time hazard models $\lambda_j(t) = \exp(\beta X_j)\lambda_0[t \exp(\beta X_j)]$. The subsequent censored data linear rank tests of H_0: $\beta = 0$ have the general form

$$v = \sum_{i=1}^{k} (X_{(i)}c_i + S^*_{(i)}C_i),$$

where the covariates are fixed, there are no tied failure times, $X_{(i)}$ is the covariate of the individual observed to fail at $t_{(i)}$, $S^*_{(i)}$ is the sum of the covariates of those censored during $(t_{(i)}, t_{(i+1)})$, and c_i and C_i are as defined by Prentice (1978). Prentice and Marek (1979) showed that in the two-sample problem v is of the form (1) with $w(t_{(i)}) = c_i - C_i$ and $b(t) \equiv 1$ whenever

$$n(t_{(i)})C_{i-1} = c_i + [n(t_{(i)}) - 1]C_i. \tag{6}$$

The substitution

$$n^{-1}(t_{(j)}) \sum_{i=j}^{k} (x_{(i)} + S^*_{(i)}) = \bar{Z}(t_{(j)})$$

in their equation (6) shows that v is equivalent to (1) with the same $w(t_{(i)})$ and $b(t_{(i)}) \equiv 1$ in the general fixed covariate problem whenever (6) holds. Mehrotra, Michalek, and Mihalko (1982) showed that (6) is always true. Hence, Prentice's linear rank tests of H_0: $\beta = 0$ can be written as $T(c_i - C_i, X)$.

Much attention has been focused on procedures that test the equality of s survival distributions. Tarone and Ware (1977) introduced a class of statistics in which the log-rank and Gehan tests for the two-sample problem and the log-rank and Breslow tests for the s-sample problem differ only in the choice of weight function. Subsequently, many authors have studied such general classes for two or more samples, including Aalen (1978), Gill (1980), Fleming et al. (1986), Breslow (1982), and Andersen et al. (1982). Before the relationship with the Tarone–Ware s-sample class is studied, the T statistic must be extended to the s-sample case. Let $\mathbf{X}'_j(t) = (X_{j1}(t), \ldots, X_{js}(t))$, where $X_{jm}(t) = 1$ if the jth individual is in group m at time t, and 0 otherwise. The obvious extensions of (1) and (4) are a $\mathbf{T}$ statistic vector whose mth component is

$$T_m(w, \mathbf{X}) = \sum_{i=1}^{k} w(t_{(i)}) \sum_{j=1}^{n} J_j(t_{(i)})[X_{jm}(t_{(i)}) - \bar{X}_m(t_{(i)})], \tag{7}$$

where $\bar{X}_m(t) = n^{-1}(t) \sum Y_j(t) X_{jm}(t)$ is the average mth component of the covariate vector and a covariance matrix $\mathbf{V}(w, \mathbf{X})$ whose (m, l)th element is

$$V_{m,l}(w, \mathbf{X}) = \sum_{i=1}^{k} w^2(t_{(i)})b^*(t_{(i)})C_{X_m,X_l}(t_{(i)}), \tag{8}$$

where $C_{X_m,X_l}(t)$ is the sample covariance, using $n^{-1}(t)$ rather than $[n(t) - 1]^{-1}$, of the mth and lth components of the covariates at risk at time t. Let $O_m(t) = \sum J_j(t)X_{jm}(t)$ be the number of observed failures in group m at time t, and $f_m(t) = \sum Y_j(t)X_{jm}(t)$ be the number at risk in group m at t. Defining $E_m(t) = f_m(t)b(t)/n(t)$ to be the hypergeometric mean for $O_m(t)$, substitution into (7) yields

$$T_m(w, \mathbf{X}) = \sum_{i=1}^{k} w(t_{(i)})[O_m(t_{(i)}) - E_m(t_{(i)})]. \tag{9}$$

Direct substitution shows that

$$C_{X_m,X_l}(t) = n^{-1}(t) \sum_{j=1}^{n} Y_j(t)[X_{jm}(t) - \bar{X}_{.m}(t)][X_{jl}(t) - \bar{X}_{.l}(t)]$$

$$= \frac{f_m(t)}{n(t)} \left(\delta_{ml} - \frac{f_l(t)}{n(t)} \right),$$

so that (8) becomes the familiar form

$$V_{m,l}(w, \mathbf{X}) = \sum_{i=1}^{k} w^2(t_{(i)}) b^*(t_{(i)}) \frac{f_m(t)[\delta_{ml} n(t) - f_l(t)]}{n^2(t)}. \tag{10}$$

Letting $\mathbf{T}' = (T_1, \ldots, T_s)$ and $\mathbf{V}_\mathbf{T}^-$ be a generalized inverse, the quadratic form $\mathbf{T}' \mathbf{V}_\mathbf{T}^- \mathbf{T}$ is identical to that of Tarone and Ware (1977) and basically the same as that of Andersen et al. (1982). The latter first introduce a separate weight function for each component but later note that most examples of interest are covered by a common weight; our w is equivalent to their L. The only other difference results from their assumption of no tied failure times so that their variance estimation is given by (10) without the tied failure time adjustment factor $[n(t) - b(t)]/[n(t) - 1]$. Because of this equivalence to the Tarone–Ware class, much is known about T in the s-sample setting. For the two-sample problem in particular, T encompasses the Efron (1967) test with

$$w(t) = n(t) f_1^{-1}(t) f_2^{-1}(t) \hat{S}_{\mathrm{KM},1}(t) \hat{S}_{\mathrm{KM},2}(t)$$

whenever $f_1(t) f_2(t) > 0$, where $\hat{S}_{\mathrm{KM},i}(t)$ is the Kaplan–Meier estimator at time t for sample i ($i = 1, 2$); it also encompasses the difference in cumulative hazards for the two samples with $w(t) = n(t) f_1^{-1}(t) f_2^{-1}(t)$ whenever $f_1(t) f_2(t) > 0$. Table 2 contains other special cases.

3.2 *Grouped Covariate*

Using the grouped covariate label defined in Section 2, it is straightforward to show that T^2/V is identical to the trend statistic first proposed by Tarone (1975) in the case $w \equiv 1$ and later extended to more general weight functions by Tarone and Ware (1977).

3.3 *Covariate Rank*

For $j \in R(t)$ let $r_j(t)$ be the rank of $X_j(t)$ among $\{X_l(t): l \in R(t)\}$, where ties are handled by midranks, and let $Z_j(t) = r_j(t)/n(t)$. Then $\bar{Z}(t) = [1 + n^{-1}(t)]/2$. Suppose there are $e(t)$ distinct covariates at time t. Let $m_1(t)$ be the number of covariates at risk equal to the smallest distinct value, $m_2(t)$ the number equal to the next smallest value, and so on. Therefore, $m_i(t)$ of the $\{r_j(t): j \in R(t)\}$ are equal to the average of $m_1(t) + \cdots + m_{i-1}(t) + 1, \ldots, m_1(t) + \cdots + m_{i-1}(t) + m_i(t)$. Note also that $n(t) = m_1(t) + \cdots + m_{e(t)}(t)$. From Lehmann (1975, p. 330, eq. A.18),

$$C_{r,r}(t) = \frac{n^2(t) - 1}{12} - \sum_{i=1}^{e(t)} \frac{m_i(t)[m_i^2(t) - 1]}{12 n(t)},$$

so that

$$C_{ZZ}(t) = \frac{1}{12} \left\{ 1 - \sum_{i=1}^{e(t)} \left[\frac{m_i(t)}{n(t)} \right]^3 \right\}. \tag{11}$$

The general covariate rank versions of T and V are obtained by substituting $[1 + n^{-1}(t)]/2$ for $\bar{Z}(t)$ and (11) into equations (1) and (4). Note in particular that $T[n(t), r(t)/n(t)] = T[1, r(t)]$. We prefer $r_j(t)/n(t)$ over $r_j(t)$ as the label since r_j ranges between 1 and $n(t)$,

Table 2
Special cases of T

Label	Setting	Weight	Equivalent test
Covariate	General	1	Cox (1972) score test of H_0: $\beta = 0$
	Fixed covariate	$c_i - C_i$	Prentice (1978) linear rank statistics
	s samples	$w(t)$	Tarone and Ware (1977); Andersen et al. (1982)
		1	Log-rank
		$n(t)$	Breslow (1970)
		$c_i - C_i$	Prentice (1978)
	2 samples	$w(t)$	Tarone–Ware (1977); Gill's (1980) K class
		1	Log-rank (Mantel, 1966)
		$n(t)$	Gehan (1965)
		$\hat{w}(t; 1, 0)$	Peto and Peto (1972)
		$\hat{w}(t; \alpha, \beta)$	Fleming et al. (1986)
	2 samples, stratify on second covariate	$w(t)$	Stratified versions of 2-sample tests
Grouped covariate	s-sample trend	1	Tarone (1975)
		$w(t)$	Tarone and Ware (1977)
$n^{-1}(t) \times$ Ranked covariate	s-sample	$w(t)$	Same as for covariate label
	Fixed continuous covariate	$n(t)$	Brown et al. (1974) modification of Kendall rank correlation
	General	$w(t)$	$T(w, r/n)$ (§3.3)
		1	Generalized log-rank (§3.3)
		$n(t)w(t)$	Weighted Kendall (§3)
	s-sample trend	$n(t)$	Jonckheere trend test (Gehan, 1965)
Transformed ranked covariates	Continuous	1	Logit rank test (O'Brien, 1978)

which inherently weighs early failure times. On the other hand, $r_j(t)/n(t)$ always stays within (0, 1]; the only weighting comes from the weight function w.

Let us consider the covariate rank version of T for the s-sample, continuous covariate, and s-sample trend problems. Define $\mathbf{X}_j'(t) = (X_{j1}(t), \ldots, X_{js}(t))$, $O_m(t)$, $f_m(t)$, and $E_m(t)$ as in Section 3.1. Then the rank of $X_{jm}(t)$ among $\{X_{im}(t): i \in R(t)\}$ is

$$r_{jm}(t) = \begin{cases} n(t) - \dfrac{f_m(t) - 1}{2} & \text{if } j \in \text{group } m \text{ at time } t, \\ \dfrac{n(t) - f_m(t) + 1}{2} & \text{if } j \in \text{group } m \text{ at time } t. \end{cases}$$

The average mth component of $(r_{j1}(t), \ldots, r_{js}(t))$ is $\bar{r}_{\cdot m}(t) = [n(t) + 1]/2$. Noting that

$$2r_{jm}(t) = [2n(t) - f_m(t) + 1]X_{jm}(t) + [n(t) - f_m(t) + 1][1 - X_{jm}(t)],$$

it is easy to show that substitution of $Z_j(t) = r_j(t)/n(t)$ into (7) yields $T_m(w, r/n) = \frac{1}{2}T_m(w, \mathbf{X})$.

Furthermore,

$$C_{r_{\cdot m}, r_{\cdot l}}(t) = n^{-1}(t) \sum_{j=1}^{n} Y_j(t)[r_{jm}(t) - \bar{r}_{\cdot m}(t)][r_{jl}(t) - \bar{r}_{\cdot l}(t)]$$

$$= \frac{n(t)}{4} \sum_{j=1}^{n} Y_j(t)[X_{jm}(t) - \bar{X}_{\cdot m}(t)][X_{jl}(t) - \bar{X}_{\cdot l}(t)]$$

so that

$$C_{Z_{\cdot m}, Z_{\cdot l}}(t) = (\tfrac{1}{4}) C_{X_{\cdot m}, X_{\cdot l}}(t) \quad \text{and} \quad V_{ml}(w, r/n) = (\tfrac{1}{4}) V_{ml}(w, \mathbf{X}).$$

Thus,

$$\{\mathbf{T}(w, r/n)\}' \mathbf{V}^{-}(w, r/n)\{\mathbf{T}(w, r/n)\} = \{\mathbf{T}(w, \mathbf{X})\}' \mathbf{V}^{-}(w, \mathbf{X})\{\mathbf{T}(w, \mathbf{X})\}$$

so that the Tarone–Ware class for the s-sample problem can be derived from the T statistic using either the covariates or the ranks of the covariates for the label. Hence, the log-rank test can be denoted by either $\mathbf{T}(1, \mathbf{X})$ or $\mathbf{T}(1, r/n)$.

Next let us consider the rank version of T when the covariate is continuous. Extending the work of Brown et al. (1974), we shall introduce a weighted Kendall rank correlation statistic modified for right-censored data for a fixed covariate. In the absence of tied failure times the rank version of T is equivalent to a specific form of the weighted Kendall statistic. To facilitate the proof we shall derive a U-statistic version of T. Assume for the continuous covariate problem $X_j = X_j(t) = X_j(0)$ and there are no tied failure times. Define

$$U_{jl} = \begin{cases} 1 & \text{if } X_j > X_l \\ \frac{1}{2} & \text{if } X_j = X_l \\ 0 & \text{if } X_j < X_l. \end{cases}$$

If $j \in R(t)$, we have

$$r_j(t) = 1 + \sum_{l \in R(t) - \{j\}} U_{jl}. \tag{12}$$

With fixed covariates the survival data typically are represented as $\{(y_j, X_j, \delta_j), j = 1, \ldots, n\}$, where y_j is the time under study for the jth individual, and $\delta_j = 1$ if he left the study through a failure and 0 if he was censored. Let t_j be the actual survival time for the jth individual which is observed only when $\delta_j = 1$. For simplicity, reorder the indices so that $y_1 \leqslant \cdots \leqslant y_n$ and $j < l$ whenever $y_j = y_l$, $\delta_j = 1$, and $\delta_l = 0$. Then from (1) and (12) the U-statistic form is

$$T[w(t)n(t), r(t)/n(t)] = \sum_{j=1}^{n-1} w(y_j)\delta_j \left(\sum_{l \in R(y_j) - \{j\}} U_{jl} - \frac{n-j}{2} \right) \tag{13}$$

since $n(y_j) = n - j + 1$ when $\delta_j = 1$. In creating a framework for the Kendall statistic, define

$$c(u, v) = \begin{cases} 1 & \text{if } u \text{ is definitely larger than } v \\ 0 & \text{if } u = v \text{ or uncertain} \\ -1 & \text{if } u \text{ is definitely smaller than } v. \end{cases}$$

For example, if $(y_1, \delta_1) = (5, 1)$, $(y_2, \delta_2) = (4, 0)$, and $(y_3, \delta_3) = (4, 1)$, then $c(t_1, t_2) = 0$, $c(t_1, t_3) = 1$, and $c(t_2, t_3) = 1$. Since the covariates $\{X_j\}$ are all known, the "uncertainty" option of $c(X_j, X_l)$ is never used. We introduce here a weighted Kendall rank statistic

$$K(v) = 2 \sum\sum_{j<l} v_{jl} c(t_j, t_l) c(X_j, X_l). \tag{14}$$

When $v_{jl} \equiv 1$, (14) is the Kendall rank statistic introduced by Brown et al. (1974, henceforth denoted by BHK).

Proposition 1 In the absence of tied failure times and when $v_{jl} = w(y_j)$, $K(w) = -4T[w(t)n(t), r(t)/n(t)]$.

Proof Since $y_1 \leqslant \cdots \leqslant y_n$ and $j < l$ whenever $y_j = y_l$, $\delta_j = 1$, and $\delta_l = 0$, then $c(t_j, t_l) = -\delta_j$ for $j < l$. Since $c(X_j, X_l) = 2U_{jl} - 1$, (14) becomes

$$\begin{aligned} K(w) &= 2 \sum\sum_{j<l} w(y_j)(-\delta_j)(2U_{jl} - 1) \\ &= \left[-4 \sum_{j=1}^{n-1} w(y_j)\delta_j \sum_{l=j+1}^{n} U_{jl} - \sum_{j=1}^{n-1} w(y_j)\delta_j \frac{n-j}{2}\right] \\ &= -4T(w, r) = -4T[w(t)n(t), r(t)/n(t)] \end{aligned}$$

by (13) and the fact $R(y_j) - \{j\} = \{j+1, \ldots, n\}$.

When there are tied failure times, T and K are not equivalent. In essence the T statistic compares covariables of simultaneously failing individuals, whereas the K statistic does not since $c(t_j, t_l) = 0$ when $t_j = t_l$. The BHK modification of the Kendall statistic is $K(1) = -4T[n(t), r(t)/n(t)]$. As such, there are two distinct advantages to $K(1)$ belonging to the T class. First, the BHK Kendall statistic requires fixed covariates, whereas as a T statistic it is extended to time-dependent covariates. Second, $K(1)$ can be extended to having weights, i.e., as $T[w^*(t)n(t), r(t)/n(t)]$ for some function $w^*(t)$, although not to the extent of the weighted Kendall statistic (14).

Note that the Kendall statistic for the general covariate problem and the Gehan two-sample statistic are both represented by $T(n, r/n)$. This confirms the observation of BHK that the Kendall statistic applied to the two-sample problem is the Gehan statistic. Since $w(t) = n(t)$, the Gehan–Kendall statistic weights early failure times more heavily; however, since the weight depends on the censoring mechanism, there can be severe consequences as shown by Prentice and Marek (1979) in the two-sample case. Let us propose a new statistic $T(1, r/n)$ that applies equal weighting across time and that is equivalent to the log-rank test in the two-sample case. This "generalized log-rank test" will be appropriate for the proportional hazards model and yet robust to outliers in the covariable space. If one wants a test that is sensitive to hazard differences which are not constant over time, then one should use the general form $T(w, r/n)$, where the weight w is chosen appropriately. Sometimes the weighting scheme can be predetermined exactly as a function of time. Oftentimes one may know only whether early, mid, or late failure times should be weighted, in which case the aforementioned $\hat{w}(t; \alpha, \beta)$ is an excellent candidate. To avoid the problems experienced by the Gehan–Kendall statistic, the weight function should be independent of the censoring distribution.

Finally, let us consider T in the s-sample trend problem. In particular, let us study the special case in which the fixed covariate labels for the s categories are ordered. In contrast to Section 3.2, no particular trend alternative is assumed here. Let us now define a Jonckheere trend statistic based on a two-sample test. Suppose the group index ordering reflects the covariate ordering. If S_{ab} is a statistic used to test the difference between group a and group b, then a reasonable test statistic of no difference among s groups against an ordered category trend alternative is the Jonckheere statistic $\sum S_{ab}$, where summation is over all group pairings $a < b$.

Proposition 2 In the s-sample trend problem the Jonckheere trend statistic (Gehan, 1965) based on Gehan's two-sample test is equivalent to the BHK censored-data modification of

Kendall rank correlation $K(1)$. In fact, $K(1) = -2 \sum S_{ab}$, where S_{ab} is the Gehan test for groups a and b and summation is over all pairs $a < b$.

Proof Without loss of generality assume the s distinct covariate values are $\{1, \ldots, s\}$. The data can be presented as $\{(y_j, \delta_j, X_j), j = 1, \ldots, n\}$, where y_j, δ_j, and X_j are as defined before. However, for simplicity of this proof it will be convenient to consider another indexing system (a, i) whereby y_{ai} is the observed time on study for the ith person in group a. The terms δ_{ai}, X_{ai}, and t_{ai} are defined similarly. Hence, $X_{ai} = a$. The data are now represented by $\{(y_{ai}, \delta_{ai}, X_{ai}), i = 1, \ldots, f_a(0)$ and $a = 1, \ldots, s\}$, where $f_a(0)$ is the size of group a at time 0. Then

$$K(1) = \sum_i \sum_j c(t_i, t_j)c(X_i, X_j)$$

$$= \sum_{a=1}^{s} \sum_{b=1}^{s} \sum_{i=1}^{f_a} \sum_{j=1}^{f_b} c(t_{ai}, t_{bj})c(X_{ai}, X_{bj}).$$

Since $c(u, v) = -c(v, u)$, $c(X_{ai}, X_{aj}) = c(a, a) = 0$, and $c(X_{ai}, X_{bj}) = c(a, b) = -1$ for $a < b$,

$$K(1) = -2 \sum_{a<b} \sum_i \sum_j c(t_{ai}, t_{bj})$$

$$= -2 \sum_{a<b} S_{ab},$$

where S_{ab} is the Gehan test comparing groups a and b. Because the Jonckheere statistic is equivalent to the Kendall statistic, it is also equivalent to the trend statistic T by Proposition 1 when there are no tied failure times and the weights are all equal to 1.

3.4 *Transformed Ranked Covariates*

In this section we consider only continuous covariates. Section 3.3 can be generalized by employing transformations of the ranked covariate labels, i.e., the label for the jth individual at time t is $Z_j(t) = g_t[r_j(t)]$, where g_t is a monotone function. In this fashion one might replace the actual covariates by the expected or approximate order statistics under some distribution for the covariates, e.g., the normal. Except for a few distributions, expected order statistics are difficult to find. To calculate approximate order statistics, let $g_t(l) = H_t^{-1}\{(l - \frac{1}{2})/n(t)\}$, where $l = 1, \ldots, n(t)$, and H_t is a distribution function. Note that $(l - \frac{1}{2})/n(t)$ is never 0 or 1 so that $g(l, t)$ will be finite. The choice $H_t = \Phi$, the normal distribution function, will give approximate normal scores. The logit function also provides a fair approximation to the normal.

Proposition 3 The trend statistics $T(1, \mathbf{Z})/\sqrt{V(1, \mathbf{Z})}$ where $Z_j(t) = \Phi^{-1}\{[r_j(t) - \frac{1}{2}]/n(t)\}$ and $Z_j(t) = \text{logit}\{[r_j(t) - \frac{1}{2}]/n(t)\}$ are equivalent to the inverse normal and logit rank tests, respectively, proposed by O'Brien (1978).

Proof This proposition follows readily upon comparison of (1) and (4) with the corresponding equations in O'Brien (1978).

As a summary to this section, Table 2 lists some tests that are members of the general class of statistics spanned by $T(w, \mathbf{Z})$.

4. Example

Table 3 specifies the posttreatment survival times and ages at time of treatment for 28 male patients with low-grade gliomas (brain tumors) who comprised a randomized study of radiotherapy with and without CCNU, an oral nitrosourea (chemotherapy). These data are

Table 3
Survival times (days) and ages (years) for 28 male patients with low-grade gliomas

Survival time	Age	Survival time	Age
6	38.4	797+	36.9
61	59.6	954	49.0
179+	18.9	967+	26.8
236	65.5	1,050	66.9
296	67.8	1,134	30.4
370	42.1	1,141	41.6
420	66.2	1,213+	35.1
474	38.1	1,349+	22.2
535	57.3	1,506+	42.0
547+	34.5	1,517+	43.1
587+	37.0	1,624	30.9
637+	30.1	1,828+	35.7
639	49.5	1,983+	29.5
747	50.9	2,237+	57.8

Table 4
Analyses of unmodified and modified Table 3 data

Statistic	Name	Standardized test statistic: Unmodified data	Standardized test statistic: Modified data
$T(1, X)$	Cox score	3.15	1.40
$T(1, r/n)$	Generalized log-rank	2.92	2.69
$T(n, r/n)$	Kendall	3.10	2.87
$T(\hat{S}_{KM}, r/n)$	Survival-weighted T	3.11	2.87
$T\left(1, \text{logit}\left(\frac{r-\frac{1}{2}}{n}\right)\right)$	Logit rank	2.88	2.35

from the Southwest Oncology Group Protocol 7983. Data indicating CCNU treatment are omitted. In order to demonstrate the differential effects of extreme covariate values on various forms of the T statistic, a second data set is created from that of Table 3 by simply changing the age of the last patient to exit the study from 57.8 to 97.8. However unlikely this modification may be for this particular data set, it is certainly within the range of extreme covariate values found in many studies. Five different versions of the normalized T statistic were computed for the unmodified and modified data and are presented in Table 4. For the unmodified data the Cox score test yields the most significant result, followed closely by the Kendall and survival function weighted covariate rank version of T, both of which emphasize early failure times. Data sets can be found for which each of these five statistics is most significant. Analysis of the modified data set, as seen in Table 4, illustrates the considerable influence of a single extreme covariate on a covariate-based procedure, such as the Cox score test which has dropped from 3.15 to 1.40, whereas the covariate rank-based procedures remain reasonably robust.

5. Discussion

The T statistic is designed to consider only a single covariate, although some extensions are possible, as seen in the s-sample problem in Section 3.1. In a multicovariate situation one often wants to test for the significance of one particular covariate while adjusting for

the effects of the others. In a nonparametric setting a popular approach to this problem is to stratify the data based on values of the covariates to be adjusted for, form the test statistic T_i and its variance estimator V_i within each stratum, and finally form a normalized statistic $(\sum T_i)(\sum V_i)^{-1/2}$, where summation is over the number of strata. For example, in a data set consisting of three covariates, sex with two levels, age group with five levels, and city with two levels, one could stratify on the ten sex–age group combinations and then test for a survival difference between the two cities using a stratified log-rank or a stratified Peto–Peto test. Of course, the decision on how to stratify a continuous covariate is often guesswork.

For significance testing purposes we would hope that $T/\sqrt{V}$ is approximately a N(0, 1) random variable under H_0. Under regularity conditions $T/\sqrt{V}$ is in fact asymptotically normal (Jones and Crowley, Technical Report No. 88-1, Department of Preventive Medicine, University of Iowa, 1988); however, in small samples the approximation may be poor. The large-sample approximation relies on the central limit theorem. Let $K = b(t_1) + \cdots + b(t_k)$ be the total number of individuals observed to fail. As seen in (1), T is approximately normal for small K under H_0 when the $\{Z_j(t)\}$ are close to normal; but when the $\{Z_j(t)\}$ are far from normal, a good approximation may require very large K. The obvious solution for small samples is to transform the covariates to be as close to normal random variables as possible. Let us consider the continuous covariate situation. One possibility would be to replace a covariate with an approximate normal score; e.g., $Z_j(t) = \Phi^{-1}[(r_j(t) - \frac{1}{2})/n(t)]$ or $Z_j(t) = \text{logit}[(r_j(t) - \frac{1}{2})/n(t)]$ as mentioned in Section 3.4. O'Brien (1978) has shown through simulation that the logit rank test, based on the latter label, has better control over the Type I error level than the Cox score test. Normal approximation to T may be poor if the distribution of covariates is skewed or there are outliers. With fixed continuous covariates another solution would be to transform the initial collection of covariates so that they are more normally distributed—perhaps a simple log transform. Certain invariance principles should be noted here: $T/\sqrt{V}$ based on the raw covariates is invariant to the linear transform $f(X) = a + bX$; $T/\sqrt{V}$ based on log-transformed covariates is invariant to the transform $f(X) = aX^b$; and $T/\sqrt{V}$ based on ranked or transformed ranked covariates is invariant to any monotone transformation.

The general trend statistic T has several very desirable properties: (i) it handles tied failure times, (ii) it allows time-dependent covariates, (iii) its variance estimator V does not require the censoring distribution to be independent of the covariate, (iv) it allows fairly general weighting schemes, and (v) it represents a fairly general class of nonparametric test procedures for survival analysis. Its two major disadvantages derive from its nonparametric nature: its principal use is for a single covariate and there are no parameters per se to estimate and interpret. The weight function should be chosen with a particular alternative in mind. For example, if one believes that the association between the failure hazard and the label is stronger early in time rather than later, the weight function should be chosen accordingly. To avoid erroneous results due to the censoring distribution, the weight function should be chosen independent of censoring-dependent quantities, such as the number at risk.

As a general class, T contains many of the survival analysis procedures in common use (Table 2), including both linear and nonlinear rank statistics. The linear rank statistics of Prentice (1978) use the ranks of the failure times and the actual covariate values, whereas the nonlinear Kendall rank correlation of BHK (1974) uses the rank of the failure times and the successive reranking of the covariates at each failure time (§3.3). There are several benefits in belonging to such a general class. If we invent a new test, or are interested in an already existing one, and we can show that it is a member of the T class, then several extensions are immediately available. For example, in the form of (1) it can accommodate weight functions and time-dependent covariates. Several of the statistics of Table 2 were

designed with neither in mind. We can also write down (4) as its variance estimator, which can handle tied failure times and is appropriate even when the censoring distribution depends on the covariate. Conditional on the covariate we still require the failure and censoring mechanisms to be independent. As an example of such an extension, as seen in Section 3.3, the Kendall statistic of BHK (1974) can be modified to handle a weight function and time-dependent covariates. Another benefit of the T class is to extend old tests into completely different covariate settings, sometimes creating new tests, while retaining the same weight function and label. For example, the two-sample log-rank procedure, as shown in Sections 3.1 and 3.3, can be written as either $T(1, \mathbf{X})$ or $T(1, r/n)$. Extending this to the continuous covariate setting, $T(1, \mathbf{X})$ is the Cox score statistic and $T(1, r/n)$ is a new procedure, a generalized log-rank. Likewise, the entire Tarone–Ware class of two-sample tests can be generalized to the continuous label statistics $T(w, \mathbf{X})$ or $T(w, r/n)$. If there are outliers in the covariable space, then instead of using the Cox score test $T(w, \mathbf{X})$, we may opt for a more robust test $T(w, \mathbf{Z})$, where the label $Z_j(t)$ is $r_j(t)/n(t)$ or perhaps the truncated covariate, the latter retaining most of the original data. For small sample sizes or highly skewed covariate situations, we may decide to replace the covariate by a transformed rank of the covariate or by a transformation of the covariate (§3.4) to have better control over the Type I error. Another benefit of such a general class is in the study of the asymptotic properties. Given a general central limit theorem for $T/\sqrt{V}$ (Jones and Crowley, unpublished technical report, 1988), asymptotic normality for any member of the T class, specified by particular weight and label functions, can be established as a corollary to the general theorem. Finally, because so many tests differ only by the choice of weight function and label, such a unifying class promotes a better understanding of the relationships among test procedures, which in turn helps in the decision as to which procedure should be used in a given situation.

Acknowledgements

This work was done while the first author was supported by National Cancer Institute training grant T32 CA09168 and research grant CA39065 and the second author by a NIGMS grant. Data for the example in Section 4 are used with the permission of Dr Harmon Eyre, Chairman of the Brain Tumor Committee of the Southwest Oncology Group, and Dr Charles A. Coltman, Jr., Chairman of the Southwest Oncology Group.

Résumé

Tarone et Ware (1977, *Biometrika* **64,** 156–160) ont développé une classe générale de test de s échantillons pour des données de survie censurées à droite; cette classe inclut la procédure du log-rang et celle de Wilcoxon modifiée. Par la suite, de nombreux auteurs ont étudié des classes pour deux ou s échantillons en détail. Dans ce papier, on montre qu'une famille de statistiques nonparamétriques unifie l'existant et permet d'engendrer de nouvelles statistiques de test pour le problème de s (≥ 2) échantillons, le problème d'une tendance pour s échantillons et celui d'une covariable continue.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics* **6,** 701–726.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1982). Linear nonparametric tests for comparison of counting processes with applications to censored survival data. *International Statistical Review* **50,** 219–258.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large-sample study. *Annals of Statistics* **10,** 1100–1120.

Breslow, N. E. (1970). A generalized Kruskal–Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika* **57,** 579–594.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30,** 89–99.

Breslow, N. E. (1982). Comparison of survival curves. In *The Practice of Clinical Trials*, M. Buyse, M. Staquet, and R. Sylvester (eds), Chap. 2, Part 6. Oxford: Oxford University Press.

Brown, B. W., Jr., Hollander, M., and Korwar, R. M. (1974). Nonparametric tests of independence for censored data, with applications to heart transplant studies. In *Reliability and Biometry, Statistical Analysis of Lifelength*, F. Proschan and R. J. Serfling (eds), 327–354. Philadelphia: Society for Industrial and Applied Mathematics.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62,** 269–276.

Efron, B. (1967). The two-sample problem with censored data. In *Proceedings of the 5th Berkeley Symposium* **4,** 831–854. Berkeley, California: University of California Press.

Fleming, T. R., Augustini, G. A., Elcombe, S. A., and Offord, K. P. (1986). The SURVDIFF procedure. Supplemental Release, SAS User's Group International. Cary, North Carolina: SAS Institute, Inc.

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

Gill, R. D. (1980). Censoring and stochastic integrals. *Mathematical Centre Tracts* **124.** Amsterdam: Mathematisch Centrum.

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

Mantel, N. (1966). Evaluation of survival data and two new rank-order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50,** 163–170.

Mehrotra, K. G., Michalek, J. E., and Mihalko, D. (1982). A relationship between two forms of linear rank procedures for censored data. *Biometrika* **69,** 674–676.

O'Brien, P. C. (1978). A nonparametric test for association with censored data. *Biometrics* **34,** 243–250.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank-invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A* **135,** 185–206.

Prentice, R. L. (1978). Linear rank tests with right-censored data. *Biometrika* **65,** 167–179.

Prentice, R. L. and Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics* **35,** 861–867.

Tarone, R. E. (1975). Tests for trend in life table analysis. *Biometrika* **62,** 679–682.

Tarone, R. E. and Ware, J. (1977). On distribution-free test for equality of survival distributions. *Biometrika* **64,** 156–160.

Received February 1987; *revised March* 1988.