Two-Sample Repeated Significance Tests Based on the Modified Wilcoxon Statistic
Author(s): Eric Slud and L. J. Wei
Source: *Journal of the American Statistical Association*, Vol. 77, No. 380 (Dec., 1982), pp. 862-868

# Two-Sample Repeated Significance Tests Based on the Modified Wilcoxon Statistic

ERIC SLUD and L.J. WEI*

The asymptotic distribution theory of sequentially computed modified-Wilcoxon scores is developed for two-sample survival data with random staggered entry and random loss to follow-up. The asymptotic covariance indicates generally dependent modified-Wilcoxon increments, contradicting (the authors' reading of) Jones and Whitehead (1979). A repeated significance testing procedure is presented for testing the equality of two survival distributions based on the asymptotic theory. The early stopping properties of this procedure are illustrated by a prostate cancer example.

KEY WORDS: Consistent estimation; Dependent increment; Gaussian process; Logrank test.

## 1. INTRODUCTION

The problem of fixed-sample comparison of survival data from two treatment groups in the presence of arbitrary right-censoring has received a great deal of attention in recent statistical literature. Two commonly used test statistics are the logrank statistic (Mantel 1966; Peto and Peto 1972) and the modified-Wilcoxon statistic (Gilbert 1962; Gehan 1965; and Breslow 1970). The choice between these two statistics has been discussed by Tarone and Ware (1977) and other authors. Generally, when it is desired to detect treatment-group differences in survival for relatively short times on treatment, the modified-Wilcoxon statistic is an especially likely tool of analysis. Peto and Peto (1972) and Prentice (1978) have proposed other generalizations of the Wilcoxon statistic that are less sensitive to censoring pattern (cf. Prentice and Marek 1979). However, since Gehan's scoring procedure is easy to explain, is widely used (Simon and Makuch 1980), and yields a test in Tarone and Ware's (1977) class of large-sample life-table analogs of optimal contingency-table tests, we restrict attention in this article to Gehan's generalization of the Wilcoxon statistic.

In most medical trials patients enter treatment serially, and data on response to treatment become available sequentially in real time. Moreover, there can be a large ethical cost in prolonging a trial past the earliest stage when an important difference in treatment-group survival can be documented. Thus it is preferable to follow the results of the trial closely and continuously as they become available. However, only a few continuous-time sequential procedures have been proposed for testing equality of survival distributions with right-censored two-sample data. Breslow (1969) and Breslow and Haug (1972) provided sequential methods of comparing exponential survival curves. Jones and Whitehead (1979) investigated sequential logrank and modified-Wilcoxon tests. The distribution theory of their test statistics was not formally developed, but a careful reading of Jones and Whitehead (1979) and Whitehead (1978) suggests that they implicitly assert sequentially computed logrank and modified-Wilcoxon scores have approximately uncorrelated increments. Our Theorem 2 and the example in Section 5 contradict that assertion for modified-Wilcoxon statistics when patient entry is not simultaneous but staggered. Finally, Nagelkerke and Hart (1980) have given a general heuristic extension of the sequential probability ratio test using partial likelihoods. Although they obtain a sequential test for censored survival data, they do not delimit conditions of applicability.

In practice, because of reporting delays and other administrative difficulties, the repeated significance test is more feasible than continuous-time sequential procedures for large-scale medical trials (cf. Armitage 1975; Pocock 1977). Recently Tsiatis (1981) rigorously derived the asymptotic joint distribution of sequentially computed logrank statistics for use in repeated significance testing. However, his test procedure requires prior knowledge of the total number of patients. Also, he does not explain (as Jones and Whitehead 1979 do) how to use repeated logrank tests if there is loss to follow-up (withdrawal or deaths from causes not under study) during the trial.

In this article, we assume that patients arrive according to a nonhomogeneous Poisson process with unknown but high intensity. The main theoretical result established is the asymptotic (for large intensity) joint normality of sequentially computed modified-Wilcoxon scores, along with consistency of an estimator of asymptotic covariance. Although the modified-Wilcoxon statistics at different time points have correlated increments, a repeated

* Eric Slud is Assistant Professor, Department of Mathematics, University of Maryland, College Park, MD 20742. L. J. Wei is Professor, Department of Statistics, George Washington University, Washington, DC 20052. This research was performed while the authors were on leave at the Clinical and Diagnostic Trials Section, Biometry Branch, National Cancer Institute. The authors are grateful to Dr. D. Byar for providing the prostate data and to Dr. R. Milton for his computer package. The comments of Drs. D. Byar and M. Gail and of two referees greatly improved the presentation of the article. The research of the second author was also supported in part by the Air Force Office of Scientific Research Grant AFOSR-81-0166.

© Journal of the American Statistical Association
December 1982, Volume 77, Number 380
Theory and Methods Section

significance test can be performed using the above asymptotic theory in the following manner: at each time $t_1 < t_2 \cdots < t_K$, a test is carried out to give nullhypothetical probability $\alpha_i$ of stopping and rejecting at $t_i$, where $\alpha_1, \ldots, \alpha_K$ are chosen in advance so that the overall significance level $\alpha = \sum_{i=1}^{K} \alpha_i$. The test boundary is determined through the asymptotic conditional distribution of the test statistic at $t_i$ given continuation to $t_i$. Precise formulation of this repeated testing procedure is given in Section 4 and illustrated with real data in Section 5.

## 2. DEFINITIONS AND TEST STATISTIC

Suppose that experimental subjects enter a clinical trial for the comparison of two treatments, $A$ and $B$, according to a nonhomogeneous Poisson process with unknown but fixed high intensity function $\lambda_c(t)$. Let the successive entry times be denoted by $U_1, U_2, \ldots$. For mathematical convenience we assume that $\lambda_c(t) = c\lambda(t)$, where $\lambda(\cdot)$ is a nonnegative function and $c$ is a positive constant that we will allow to become large. The arrival point-process is the superposition of two independent Poisson processes with intensities $c\lambda_A(t)$ and $c\lambda_B(t)$ governing entry to treatments $A$ and $B$, respectively, so that $\lambda(t) = \lambda_A(t) + \lambda_B(t)$. We let $Z_i$ be the indicator variable equal to 1 if the arrival at time $U_i$ is assigned to group $A$. For fixed $Z_i$, the $i$th patient has independent latent times $X_i$ and $Y_i$ of survival and loss to follow-up (due to withdrawal from the trial or death from causes other than the one under study). The pairs $\{(X_i, Y_i)\}$ are assumed conditionally independent of each other and of $\{U_i\}$ given $\{Z_i\}$.

Let $F_A$ ($F_B$) denote the distribution function of the survival time under treatment $A$ ($B$), and let $G_A$ ($G_B$) be the corresponding distribution function of the time until loss to follow-up. Throughout the article we use overbars above distribution functions to denote survival functions; for example, $\bar{F}_A = 1 - F_A$. Let $N_A(t)$ and $N_B(t)$ be the numbers of group $A$ and group $B$ patients entered by time $t$, with $N(t) \equiv N_A(t) + N_B(t)$, and define $\Lambda_A(t) = \int_0^t \lambda_A(x)dx$, $\Lambda_B(t) = \int_0^t \lambda_B(x)dx$, $\Lambda(t) = \Lambda_A(t) + \Lambda_B(t)$. Then $N_A(t)$ and $N_B(t)$ are independent Poisson variables with means $c\Lambda_A(t)$ and $c\Lambda_B(t)$.

The observable data at time $t$ consist of the quantities $N_A(t)$, $N_B(t)$, $Z_i$, $T_i(t) = \min(X_i, Y_i, t - U_i)$, and $\Delta_i(t) = I[X_i \le \min(Y_i, t - U_i)]$ for $i = 1, \ldots, N(t)$, where $I[\cdot]$ denotes indicator function. The hypothesis we are interested in testing is $H_0: F_A = F_B$. The Gehan-Gilbert score function $\phi$ is defined by

$$\phi(a, \delta; b, \epsilon) = \begin{cases} 1, & \text{if } a < b, \delta = 1, \\ -1, & \text{if } a > b, \epsilon = 1, \\ 0, & \text{otherwise.} \end{cases}$$

The modified-Wilcoxon statistic at time $t$ is given by

$$W_c(t) = (N_A(t)N_B(t)N(t))^{-1/2} \sum_{i=1}^{N(t)} \sum_{j=1}^{N(t)} Z_i(1 - Z_j) \times \phi(T_i(t), \Delta_i(t); T_j(t), \Delta_j(t)),$$

whenever both $N_A(t)$ and $N_B(t) \ge 1$, and $W_c(t) = 0$ otherwise.

## 3. THE ASYMPTOTIC DISTRIBUTION OF $W_c(t)$

The asymptotic joint normality of $(W_c(t_1), \ldots, W_c(t_K))$ as $c \to \infty$ is provided for the general case $F_A \neq F_B$ by the following theorem. A closed-form expression for the limiting covariance under the null hypothesis $F_A = F_B$ is presented in Theorem 2, and the corollary gives consistent covariance-estimators. All proofs, and some technical comments regarding weak convergence of $W_c(\cdot)$ suitably standardized to a Gaussian process, are deferred to the Appendix.

*Theorem 1.* Let

$$p(t) = [\Lambda_A(t)\Lambda_B(t)]^{-1} \int_0^t \Lambda_A(t - x)\Lambda_B(t - x) \times \bar{G}_A(x)\bar{G}_B(x)(\bar{F}_B(x)dF_A(x) - \bar{F}_A(x)\,dF_B(x)),$$

let

$$\phi_{ij}(t) = \phi(T_i(t), \Delta_i(t); T_j(t), \Delta_j(t)),$$

and

$$W_c^*(t) = [N_A(t)N_B(t)N(t)]^{-1/2} \times \sum_{i=1}^{N(t)} \sum_{j=1}^{N(t)} Z_i(1 - Z_j)(\phi_{ij}(t) - p(t)).$$

Then, for fixed $0 < t_1 < t_2 < \cdots < t_K < \infty$ with $\Lambda(t_1) > 0$, as $c \to \infty$, $(W_c^*(t_1), \ldots, W_c^*(t_K))$ converges in distribution to a multivariate normal random vector.

*Remark.* It is not difficult to see that the limiting normal distribution Theorem 1 is nonsingular if $\lambda_A(t)$, $\lambda_B(t)$ are everywhere strictly positive and $t_1$ is large enough that $P(\Delta_1(t_1) = 1) > 0$.

*Theorem 2.* Under the null hypothesis $H_0$, $p(t) = 0$ for $t > 0$, and if $s \le t$, $(W_c^*(s), W_c^*(t)) = (W_c(s), W_c(t))$ converges in distribution as $c \to \infty$ to a bivariate normal with mean $\mathbf{0}$ and covariance

$$\begin{aligned} \sigma(s, t) &= \lim_{c \to \infty} E(W_c(s)W_c(t)) \\ &= [\Lambda(s)\Lambda(t)\Lambda_A(s)\Lambda_A(t)\Lambda_B(s)\Lambda_B(t)]^{-1/2} \\ &\quad \times \int_0^s \bar{F}^2(x)\Lambda_A(s - x)\Lambda_B(s - x)\bar{G}_A(x)\bar{G}_B(x) \\ &\quad \times (\Lambda_B(t - x)\bar{G}_B(x) + \Lambda_A(t - x)\bar{G}_A(x))dF(x), \end{aligned}$$

where $F = F_A = F_B$.

A consistent estimator of $\sigma(s, t)$ can be formed by sutstituting empirical estimators into the above integral (for details and further justification, see Appendix), yielding

*Corollary.* Under the null hypothesis $H_0$, as $c \to \infty$ the covariance $E(W_c(s)W_c(t))$ can be consistently estimated

*Table 1. VACURG Study 1 Stage 1 Prostatic Cancer Data*

| | | | | | |
|---|---|---|---|---|---|
| | | *Group A (Prostatectomy and Estrogen)* | | | |
| (10,84), | (18,63), | (52,143*), | (54,65), | (70,117), | (79,0), |
| (11,61), | (46,157*), | (50,77), | (65,136*), | (25,75), | (29,117), |
| (47,19), | (10,20), | (13,45), | (24,151*), | (27,30), | (30,0), |
| (32,68), | (40,140), | (59,5), | (78,108*), | (28,163*), | (66,66), |
| (81,12), | (33,0), | (48,144*), | (52,4), | (59,33), | (57,128), |
| (36,199*), | (36,55), | (38,172), | (53,6), | (57,177*), | (58,93), |
| (78,107), | (58,171), | (61,26), | (65,140*), | (69,13) | (45,37), |
| (49,14), | | | | | |
| | | *Group B (Prostatectomy and Placebo)* | | | |
| (5,84), | (45,46), | (42,112), | (51,60), | (53,142*), | (70,125*), |
| (77,119*), | (11,142), | (60,146*), | (61,76), | (62,38), | (14,89), |
| (50,45), | (26,111), | (11,178*), | (14,5), | (16,173*), | (19,89), |
| (27,133), | (30,163*), | (41,155*), | (64,28), | (76,114), | (10,32), |
| (22,166*), | (61,26), | (30,192*), | (39,155*), | (41,93), | (42,29), |
| (49,65), | (57,130), | (72,120*), | (77,103), | (30,110), | (37,98), |
| (38,95), | (41,70), | (58,156*), | (70,113), | (70,38), | (74,148*), |
| (63,131*), | (62,61), | (77,117*), | (73,56), | | |

NOTE: In each pair $(s, t)$, $s$ is the entry time in months after 1960 and $t$ is the time of survival or loss to follow-up in months. Loss to follow-up is indicated by *.

for $0 < s \leq t$ by

$$\hat{\sigma}(s, t) = [N(s)N(t)N_A(s)N_A(t)N_B(s)N_B(t)]^{-1/2} \times \sum_{i=1}^{N(t)} \sum_{j=1}^{N(s)} \sum_{k=1}^{N(s)} \Delta_i(t) Z_j (1 - Z_k) \times I[T_i(t) \leq \min(T_j(s), T_k(s))].$$

*Remark.* For fixed $t$, and arbitrary $i = 1, \ldots, N(t)$ and $0 \leq u \leq t$, let

$$\rho_i(u) \equiv \sum_{k=1}^{N(u)} Z_k I[X_k(u) \geq X_i(t)],$$

$$r_i(u) \equiv \sum_{k=1}^{N(u)} I[X_k(u) \geq X_i(t)].$$

Then the triple sum in the expression for $\hat{\sigma}(s, t)$ is simply

$$\sum_{i:0 \leq T_i(t) \leq s} \Delta_i(t) \rho_i(s) (r_i(s) - \rho_i(s)).$$

This formula for $\hat{\sigma}(s, t)$ in terms of rank statistics shows that $\hat{\sigma}(s, s)$ agrees precisely with the variance estimator that Tarone and Ware (1977) proposed by analogy with Mantel-Haenszel denominators.

## 4. THE REPEATED SIGNIFICANCE TEST

Suppose that significance tests based on the modified-Wilcoxon statistic are to be performed at time points $t_1 < t_2 < \cdots < t_K$, proceeding to $t_K$ only if $H_0$ has not previously been rejected. The significance levels $\alpha_1, \ldots, \alpha_K$ must be prespecified as in Section 1 so that the overall significance level $\alpha = \sum_{i=1}^{K} \alpha_i$ is fixed. At time point $t_l$, $1 \leq l \leq K$, the cutoff or boundary point $d_l$ can be determined as follows: let $(V_1, \ldots, V_l)$ be multivariate normal with mean $\mathbf{0}$ and variance-covariance matrix $\Sigma_l = (\sigma_{ij})$, where $\sigma_{ij} = \hat{\sigma}(t_i, t_j)/(\hat{\sigma}(t_i, t_i)\hat{\sigma}(t_j, t_j))^{1/2}$, $1 \leq i \leq j \leq l$. Then

$$P(|V_1| < d_1, \ldots, |V_{l-1}| < d_{l-1}, |V_l| \geq d_l) = \alpha_l. \tag{4.1}$$

If the observed Wilcoxon score $|W_c(t_l)| \geq d_l[\hat{\sigma}(t_l, t_l)]^{1/2}$, we reject the null hypothesis $H_0$ at $t_l$. For future reference, we also define the $p$ value at $t_l$ in the event of no rejection of $H_0$ before $t_l$ by

$$p_l = P(|V_1| < d_1, \ldots, |V_{l-1}| < d_{l-1}, |V_l| \geq W_c(t_l)/[\hat{\sigma}(t_l, t_l)]^{-1/2}).$$

*Table 2. Summary of VACURG Study 1 Stage 1 Data*

| | | *Time Intervals (years)* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *0–3* | *3–5* | *5–6* | *6–9* | *9–10* | *10–12* | *12–15* | *15–20* | *Total* |
| Group A (estrogen) | Entrants | 15 | 18 | 6 | 4 | 0 | 0 | 0 | 0 | 43 |
| | Deaths | 3 | 4 | 4 | 11 | 1 | 2 | 3 | 3 | 31 |
| Group B (placebo) | Entrants | 14 | 16 | 10 | 6 | 0 | 0 | 0 | 0 | 46 |
| | Deaths | 1 | 1 | 1 | 9 | 3 | 7 | 4 | 3 | 29 |
| 20-Year Summary Statistics (normal deviates): | | | | | | | | | | |
| | | Mantel-Haenszel = 1.442, p = .149 | | | | | | | | |
| | | Gehan-Wilcoxon = 1.946, p = .052 | | | | | | | | |

*Table 3. Estimated Covariance Matrices and Standardized Modified-Wilcoxon Values*

| | t (years) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 3 | 6 | 9 | 12 | 5 | 10 | 15 | 20 |
| $\hat{\sigma}(s,t)$ | .0959 | .0391 | .0420 | .0420 | .0700 | .0777 | .0777 | .0777 |
| Matrices | .0391 | .0862 | .1045 | .1046 | .0777 | .2262 | .2350 | .2350 |
| | .0420 | .1045 | .2019 | .2104 | .0777 | .2350 | .3047 | .3082 |
| | .0420 | .1046 | .2104 | .2763 | .0777 | .2350 | .3082 | .3146 |
| $W_c(t)/(\hat{\sigma}(t,t))^{1/2}$ | .9931 | 2.299 | 2.780 | 2.312 | 1.740 | 2.611 | 2.068 | 1.946 |

To evaluate (4.1), multivariate normal integrals are required. Several numerical methods are available for calculating these (cf. Johnson and Kotz 1972, p. 43). Our computations for this article were made with the multivariate normal integral package of Milton (1972), which he kindly supplied us.

In practice, the necessity of computing $l$-dimensional integrals to find the boundary values $d_l$ and $p$ values $p_l$ may deter casual or experimental application of the method, although the calculations are quite feasible (say for $K \leq 8$, using Milton's 1972 program) both for the design and the analysis of real clinical trials. A simple but effective approximate technique of calculation should facilitate experimentation with these repeated significance tests, namely: when $\sigma_{ij} \cdot \sigma_{jk}$ is nearly equal to $\sigma_{ik}$ for all $1 \leq i \leq j \leq k \leq l$, $(V_1, \ldots, V_l)$ can be treated as a Gaussian Markov sequence, and $\alpha_l/(1 - \sum_{i=1}^{l-1} \alpha_i)$ is approximated by the bivariate-normal conditional probability $P(|V_l| \geq d_l \,||\, V_{l-1}| < d_{l-1})$. This approximation was quite accurate for calculations in the example of Section 5, yielding errors in $d_l$ and $p_l$ uniformly less than .035 and .0025, respectively.

## 5. AN EXAMPLE

The data in Table 1 are taken from Study 1 (of Stage I prostatic cancer) conducted by the Veterans Administration Cooperative Urological Research Group (VACURG). The study is explained and the data are provisionally displayed in Byar (1972). Patients were admitted between 1960 and 1967 and randomly allocated to one or two treatment groups. Group $A$ patients were treated by radical prostatectomy and 5.0 mg estrogen (DES) daily by mouth, while group $B$ had prostatectomy followed by daily oral placebo. Out of 120 patients, 31 were later excluded from analysis for failure to adhere to protocol. Of the 29 patients not known to have died, all but two (one from each group) were still alive and in the study at the end of the 15th year (1974). The times on study and times of entry have been grouped by months and reported in Table 2 by years from January 1, 1960.

The summary Mantel-Haenszel and Gehan-Wilcoxon (two-sided) $p$ values differ mainly because of excess deaths in Group $A$, which occurred both chronologically early and at times-on-study of less than two years. Therefore it is of interest to compute sequential modified-Wilcoxon values and observe the behavior of our repeated significance test on this set of data. Table 3 shows the modified-Wilcoxon scores (computed from data grouped by months) and estimated variance-covariance matrices $\hat{\sigma}(s, t)$.

The $4 \times 4$ $\hat{\sigma}$ matrices were used in a repeated two-sided significance test with four looks, at times 3, 6, 9, 12 or 5, 10, 15, 20 (years after 1960). In accordance with our recommendations in Section 6, we chose significance levels $\alpha_i$ increasing in order to achieve a two-sided stopping boundary narrowing as time increases. The boundary values and $p$ values are displayed in Table 4 for two such sets of $\alpha_i$. Rejection of the null hypothesis (and early stopping) would have occurred in 1969 if analysis were performed every three years and in 1970 if every 5. It is interesting to note that the final boundary values (at time 12 or 20) would not have been inordinately larger than the fixed-sample cut-off 1.96.

The bivariate-normal computations for this example were performed with IMSL routine MDBNOR; the mul-

*Table 4. Repeated Significance Test Boundaries and p Values[a]: $\alpha = .05$*

| | $t_i$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 3 | 6 | 9 | 12 | 5 | 10 | 15 | 20 |
| $\alpha_i$ | .0075 | .0125 | .0150 | .015 | .0075 | .0125 | .0150 | .015 |
| Boundary | 2.674 | 2.478 | 2.307 | 2.162 | 2.674 | 2.453 | 2.240 | 2.041 |
| $p$ value | .321 | .020 | .008 | .008 | .082 | .008 | .027 | .024 |
| $\alpha_i$ | .0050 | .0100 | .0150 | .020 | .0050 | .0100 | .0150 | .020 |
| Boundary | 2.807 | 2.560 | 2.325 | 2.095 | 2.807 | 2.540 | 2.272 | 2.002 |
| $p$ value | .321 | .021 | .003 | .009 | .082 | .008 | .029 | .026 |

[a] As defined in Section 4.

tivariate normal integrals of dimensions three and four were computed with Milton's (1972) Fortran subroutine. All triple integrals (for significance levels and $p$ values) are accurate to within .0001; the quadruple integrals are accurate to within .001 for the five-year between looks cases, and to within .0025 for the three-year cases.

## 6. DISCUSSION

The treatment of purely sequential modified-Wilcoxon tests given by Jones and Whitehead (1979) is correct, as their simulations show, if all patients enter simultaneously. It can be seen from our Theorem 2 that $\sigma(s, s) = \sigma(s, t)$ if all the $U_i$ take the value 0 with probability 1. However, if patient entry is staggered in time, it is impossible to provide a test boundary of predetermined shape (for $W_c(t)/\hat{\sigma}(t, t)^{1/2}$) that achieves a preset overall significance level $\alpha$. Note that this objection applies generally to repeated significance tests even in the case of independent increments since variance increments cannot ordinarily be assumed equal. For example, to use the repeated logrank test of Tsiatis (1981) or a repeated test using the Prentice (1978) or Peto and Peto (1972) generalized Wilcoxon, developed in Tsiatis (1982), the stopping boundary must depend on the data through the estimated variance increments. In all such situations, the approach of our Section 4 yields easily constructed data-dependent boundaries.

Further insight into the nature of correlation between modified-Wilcoxon increments can be gained from the final result of the Appendix. The generality of the assumption of that lemma suggests (at least when arrival times follow approximately homogeneous Poisson processes) that under $H_0$ strictly negatively correlated increments are the typical case. The lemma applies (but is not restricted) to cases where the underlying survival density is nonincreasing on $[0, \infty)$. Moreover, the assumption becomes easier to satisfy in the presence of an increasing hazard from competing causes of death. The importance of the lemma lies in pointing up the undesirability for repeated modified-Wilcoxon tests of stopping boundaries with $d_i$ nearly equal. If $W_c(t_i)/(\hat{\sigma}(t_i, t_i))^{1/2} > d_i$ (but is not too much greater), then negatively correlated increments and $d_{i+1} \geq d_i$ would imply under the null hypothesis that $W_c(t_{i+1})/(\hat{\sigma}(t_{i+1}, t_{i+1}))^{1/2}$ is likely to be less than $d_{i+1}$. If for some reason a decision to stop a trial based on $W_c(t_i)$ did not take effect until $t_{i+1}$, an upsetting and paradoxical situation could arise in practice, since a new analysis at $t_{i+1}$ is likely to show a nonsignificant result. For this reason, we recommend that our repeated significance-testing procedure be used only with a steadily increasing sequence of $\alpha_i$ (to achieve a fixed $\alpha$ with steadily decreasing cutoffs $d_i$). This recommendation is strengthened by the practical desire—tempered by considerations of ethical costs versus maximum power, which are particular to each clinical trial—that the final cutoff $d_K$ not be too much larger than the fixed-sample cutoff $\Phi^{-1}(1 - \alpha/2)$.

The problem of properly choosing sequences $\alpha_1, \ldots, \alpha_K$ and $t_1, \ldots, t_K$ for the design and analysis of sequential trials is currently being investigated. Two possible methods of prescribing $\alpha_i$ are given by

$$\alpha_i = P(|V_1| < d'_1, \ldots, |V_{i-1}| < d'_{i-1}, |V_i| \geq d'_i),$$

where $(V_1, \ldots, V_K)$ is multivariate normal with mean 0 and $\operatorname{cov}(V_i, V_j) = \min(i, j)/(i, j)^{1/2}$, and $d'_i$ can be taken constant corresponding to the boundary of Armitage (1975) and Pocock (1977), or according to a Wald sequential probability ratio test (SPRT) boundary.

## APPENDIX

### A.1 Proof of Theorem 1

For fixed $N_A(t), N_B(t)$, the entry times of group $A$ and $B$ patients up to time $t$ are random samples from the distribution functions $\Lambda_A(\cdot)/\Lambda_A(t)$ and $\Lambda_B(\cdot)/\Lambda_B(t)$, respectively, on $[0, t]$. Hence the random variable $W_c^*(t)$ can be viewed as a $U$ statistic (cf. Hoeffding 1948). If we let $W_c^{**}(t)$ be the projection (Lehmann 1975, p. 362) of $W_c^*(t)$, then it follows from a similar argument provided by Wei (1980) that

$$E((W_c^*(t) - W_c^{**}(t))^2 \mid N_A(t), N_B(t)) \leq 2/(N(t) + 1).$$

Therefore if $\Lambda(t) > 0$, as $c \to \infty$, $E(W_c^*(t) - W_c^{**}(t))^2 \to 0$. By Corollary 6 of Lehmann (1975, p. 389), asymptotically the vector $(W_c^*(t_1), \ldots, W_c^*(t_K))$ has the same distribution as $(W_c^{**}(t_1), \ldots, W_c^{**}(t_K))$. By the Weak Law of Large Numbers, as $c \to \infty$, the vectors $(N_A(t_1)/c, \ldots, N_A(t_K)/c)$ and $(N_B(t_1)/c, \ldots, N_B(t_K)/c)$ converge in probability to constant vectors. Now the Cramér-Wold device (Billingsley 1968, Theorem 7.7) and Billingsley's Theorem 17.1 imply the asymptotic joint normality of $(W_c^{**}(t_1), \ldots, W_c^{**}(t_K))$ and hence of

$$(W_c^*(t_1), \ldots, W_c^*(t_K)).$$

### A.2 Weak Convergence of $W_c^*(\cdot)$

The method of proving weak convergence of $W_c^*(\cdot)$ to a Gaussian process is to verify Billingsley's (1968, 15.44) criterion with $\alpha = 1$, $\gamma = 2$ in Theorem 15.7, proving tightness of $\{W_c^*(\cdot): c \geq 1\}$ in $D[T_0, T]$ whenever $\Lambda_A(T_0)\Lambda_B(T_0) > 0$ and $T > T_0$. The derivation of an upper bound on $E((W_c^*(t) - W_c^*(s))^2(W_c^*(u) - W_c^*(t))^2)$ required to prove (15.44) is straightforward but lengthy, and we omit it. Together with the weak convergence of finite-dimensional distributions proved in Theorem 1, tightness implies weak convergence to a Gaussian process in $D[T_0, T]$.

### A.3 Proof of Theorem 2

For group $\gamma$ ($\gamma = A, B$) given $N_\gamma(t)$, let $\{(X_{\gamma i}, Y_{\gamma i}, U_{\gamma i}): i = 1, \ldots, N_\gamma(t)\}$ denote the randomly permuted set of triples $(X_j, Y_j, U_j)$ corresponding to arrivals of group $\gamma$ patients up to time $t$. Then let $T_{\gamma i}(u) = \min(X_{\gamma i}, Y_{\gamma i}, U_{\gamma i})$, $\Delta_{\gamma i}(u) = I[X_{\gamma i} = T_{\gamma i}(u)]$ for $u = s, t$, where

$s < t$ is fixed. Note that given $N_\gamma(t)$, the triples $(X_{\gamma i}, Y_{\gamma i}, U_{\gamma i})$ are independent and identically distributed.

Now

$$E(W_c(s)W_c(t) \mid N_\gamma(s), N_\gamma(t), \gamma = A, B) = [N_A(s)N_B(s)N_A(t)N_B(t)N(s)N(t)]^{-1/2} \times \sum_{i=1}^{N_A(t)} \sum_{j=1}^{N_B(t)} \sum_{k=1}^{N_A(t)} \sum_{l=1}^{N_B(t)} g(i, j, k, l),$$

where

$$g(i, j, k, l) = E[\varphi(T_{Ai}(s), \Delta_{Ai}(s); T_{Bj}(s), \Delta_{Bj}(s)) \times \varphi(T_{Ak}(t), \Delta_{Ak}(t); T_{Bl}(t), \Delta_{Bl}(t)) \mid N_\gamma(s), N_\gamma(t), \gamma = A, B].$$

Note that the variable whose expectation defines $g(i, j, k, l)$ can be nonzero only when $U_{Ai}, U_{Bj} \le s$.

If $i \neq k$ and $j \neq l$, under the null hypothesis $g(i, j, k, l) = 0$. On the other hand, the number of terms for which $i = k$ and $j = l$ is $N_A(t)N_B(t)$, and since $g$ is bounded by 1, as $c \to \infty$ these terms are asymptotically negligible in the conditional expectation of $W_c(s)W_c(t)$.

To treat the terms with $i = k, j \neq l$, we apply the definition of the score function $\phi$ to obtain

$$g(i, j, k, l) = \{P[X_{Ai} < \min(X_{Bj}, X_{Bl}, R) \text{ or } \max(X_{Bj}, X_{Bl}) < \min(X_{Ai}, R)] - P[X_{Bl} < X_{Ai} < \min(X_{Bj}, R) \text{ or } X_{Bj} < X_{Ai} < \min(X_{Bl}, R)]\},$$

where we have written $R = \min(Y_{Ai}, Y_{Bj}, Y_{Bl}, s - U_{Ai}, s - U_{Bj}, t - U_{Bl})$. Using the conditional exchangeability given $R$ of $X_{Ai}, X_{Bj}, X_{Bl}$ (in fact, independence with common distribution function $F_A = F_B = F$ under $H_0$), we conclude

$$g(i, j, k, l) = \left[\int_0^s \bar{F}^2(x)\bar{G}_A(x)\bar{G}_B{}^2(x) \frac{\Lambda_A(s - x)}{\Lambda_A(s)} \frac{\Lambda_B(s - x)}{\Lambda_B(s)} \frac{\Lambda_B(t - x)}{\Lambda_B(t)} dF(x)\right] \times \frac{N_A(s)}{N_A(t)} \frac{N_B(s)}{N_B(t)}$$

and there are $N_A(t)N_B(t)(N_B(t) - 1)$ such terms. Similarly, calculating the terms $g(i, j, k, j)$ with $i \neq k$, then applying the convergence in probability of $N_A(\cdot)/c, N_B(\cdot)/c$ to $\Lambda_A(\cdot), \Lambda_B(\cdot)$ as $c \to \infty$, and taking the expectation of the conditional expectation of $W_c(s)W_c(t)$, finishes the proof of Theorem 2.

## A.4 Proof of Corollary

Using the notations of the previous proof, we can express the covariance

$$\sigma(s, t) = \left[\frac{\Lambda_A(s)\Lambda_B(s)}{\Lambda_A(t)\Lambda_B(t)\Lambda(s)\Lambda(t)}\right]^{1/2} \times \{\Lambda_B(t) \times P[\Delta_{A1}(t) = 1, T_{A1}(t) \le \min(s, T_{B1}(s), T_{A2}(s))] + \Lambda_A(t) \times P[\Delta_{B1}(t) = 1, T_{B1}(t) \le \min(s, T_{A1}(s), T_{B2}(s))]\}.$$

Substituting the consistent estimates $N_A(\cdot)/c, N_B(\cdot)/c$ for $\Lambda_A(\cdot)$ and $\Lambda_B(\cdot)$, and replacing the bracketed probabilities by their empirical estimators, gives the form $\hat{\sigma}(s, t)$ as stated in our corollary.

## A.5 Lemma: Condition for Negatively Correlated Increments of $W_c(\cdot)$ Under $H_0$

Under the null hypothesis $H_0$, if $\lambda_A(\cdot) \equiv \lambda_A$ and $\lambda(\cdot) \equiv \lambda$ are constant with $\lambda_A/\lambda = \eta$, and if

$$\int_0^t \bar{F}_A{}^2(x)\bar{G}_A(x)\bar{G}_B(x)(\eta\bar{G}_A(x) + (1 - \eta)\bar{G}_B(x))dF_A(x)$$

is concave in $t$, then for all $0 < t < u$, $\sigma(t, t) > \sigma(t, u)$.

The proof follows from the formula from Theorem 2, substituting $\lambda_A(\cdot)/\lambda(\cdot) \equiv \eta$:

$$\sigma(t, u) - \sigma(t, t) = \int_0^t \bar{F}_A{}^2(x)\bar{G}_A(x)\bar{G}_B(x)(\eta\bar{G}_A(x) + (1 - \eta)\bar{G}_B(x))\left[\frac{\Lambda(t - x)}{\Lambda(t)}\right]^2 \times \left[\left(\frac{\Lambda(t)}{\Lambda(u)}\right)^{1/2} \frac{\Lambda(u - x)}{\Lambda(u)} - \frac{\Lambda(t - x)}{\Lambda(t)}\right] dF_A(x).$$

[*Received August 1981. Revised February 1982.*]

## REFERENCES


ARMITAGE, P. (1975), *Sequential Medical Trials* (2nd ed.), Oxford: Blackwell.

BILLINGSLEY, P. (1968), *Weak Convergence of Probability Measures*, New York: John Wiley.

BRESLOW, N. (1969), "On Large Sample Sequential Analysis With Applications to Survivorship Data," *Journal of Applied Probability*, 6, 261–274.

——— (1970), "A Generalized Kruskal-Wallis Test for Comparing $K$ Samples Subject to Unequal Patterns of Censorship," *Biometrika*, 57, 579–594.

BRESLOW, N., and HAUG, C. (1972), "Sequential Comparison of Exponential Survival Curves," *Journal of the American Statistical Association*, 67, 691–697.

BYAR, D.P. (1972), "Treatment of Prostatic Cancer: Studies by the Veterans Administration Cooperative Urological Research Group," *Bulletin of the New York Academy of Medicine*, 48, 751–766.

GEHAN, E.A. (1965), "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," *Biometrika*, 52, 203–223.

GILBERT, J.P. (1962), "Random Censorship," unpublished Ph.D. dissertation, University of Chicago. Statistics Department.

HOEFFDING, W. (1948), "A Class of Statistics With Asymptotically Normal Distribution," *Annals of Mathematical Statistics*, 19, 293.

JOHNSON, N.L., and KOTZ, S. (1972), *Distributions in Statistics; Continuous Multivariate Distributions*, New York: John Wiley.

JONES, D., and WHITEHEAD, J. (1979), "Sequential Forms of the Log Rank and Modified Wilcoxon Tests for Censored Data," *Biometrika*, 66, 105–113.

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.

MILTON, R.C. (1972), "Computer Evaluation of the Multivariate Normal Integral," *Technometrics*, 14, 881–889.
NAGELKERKE, N.J.D., and HART, A.A.M. (1980), "The Sequential Comparison of Survival Curves," *Biometrika*, 67, 247–249.
PETO, R., and PETO, J. (1972), "Asymptotically Efficient Rank Invariant Test Procedures," *Journal of the Royal Statistical Society*, Ser. A, 135, 185–206.
POCOCK, S.J. (1977), "Group Sequential Methods in the Design and Analysis of Clinical Trials," *Biometrika*, 64, 191–200.
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.
SIMON, R.M., and MAKUCH, R.W. (1980), "Letter to the Editor [on the article of Prentice and Marek cited above]," *Biometrics*, 36, 354.
TARONE, R.E., and WARE, J. (1977), "On Distribution-Free Tests for Equality of Survival Distributions," *Biometrika*, 64, 156–160.
TSIATIS, A.A. (1981), "The Asymptotic Joint Distribution of the Efficient Scores Test for the Proportional Hazards Model Calculated Over Time," *Biometrika*, 68, 311–315.
——— (1982), "Repeated Significance Testing for a General Class of Statistics Used in Censored Survival Analysis," *Journal of the American Statistical Association*, 77, 855–861.
WEI, L.J. (1980), "A Generalized Gehan and Gilbert Test for Paired Observations That Are Subject to Arbitrary Right Censorship," *Journal of the American Statistical Association*, 75, 634–637.
WHITEHEAD, J. (1978), "Large Sample Sequential Methods With Application to the Analysis of 2 × 2 Contingency Tables," *Biometrika*, 65, 351–356.