Nonparametric Sequential Testing in Clinical Trials with Incomplete Multivariate Observations

D. Y. Lin

Biometrika, Vol. 78, No. 1. (Mar., 1991), pp. 123-131.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28199103%2978%3A1%3C123%3ANSTICT%3E2.0.CO%3B2-B

Biometrika is currently published by Biometrika Trust.

Biometrika (1991), **78**, 1, *pp.* 123-31
Printed in Great Britain

Nonparametric sequential testing in clinical trials with incomplete multivariate observations

BY D. Y. LIN

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

SUMMARY

This paper addresses sequential testing in randomized clinical trials with multiple endpoints. Patients enter treatments serially and are subject to random loss to followup. The endpoints of interest may be time-to-event variables or other quantitative measurements. The proposed test statistic at a given look is a weighted sum of the linear rank statistics with respect to the marginal distributions of the multiple endpoints. The weights can be chosen to maximize asymptotic power against certain local alternatives. Stopping boundaries are obtained from the asymptotic joint distribution of the proposed test statistics calculated at different looks. This new approach preserves a single preset overall significance level and can lead to quicker termination of the trial than sequential procedures based on single endpoints. An example taken from an AIDS clinical trial is presented.

Some key words: Asymptotic joint distribution; Group sequential design; Interim analysis; Linear rank statistic; Martingale; Multiple endpoints; Repeated significance testing.

1. INTRODUCTION

In many long-term clinical trials, patients are enrolled serially, and the data on several endpoints or outcome measures become available sequentially in calendar time. It is customary to monitor these data periodically so that the study can be terminated as soon as the evidence of important differences between treatment arms is found. The two problems of the multiplicity of data simultaneously arising here, namely sequential significance testing and multiple endpoints, have conventionally been dealt with separately in statistical publications. The current literature on sequential statistical methods is fairly extensive; see Jennison & Turnbull (1989) for an excellent survey in this area. In contrast, there have been only a few papers on the issue of multiple endpoints in clinical trials (O'Brien, 1984; Wei & Johnson, 1985; Pocock, Geller & Tsiatis, 1987).

Recent years have been seen a growing interest in combining separate solutions to the problems of sequential testing and multiple endpoints. There are at least two reasons for such pursuits. First, it is often desirable to make a single overall probability statement on whether the experimental therapy is efficacious with respect to all important endpoints over the course of the study. This need is particularly acute when the emphasis of one endpoint over the rest is judgmental. Secondly, effective sample sizes are usually small in early looks of a sequential trial so that, when separate tests on each endpoint fail to demonstrate statistical significance, combined evidence from several outcome measures may be convincing.

Some parametric methods for the interim analysis of multiple endpoints have appeared in the literature. For example, Geary (1988) presented a sequential testing procedure for

normally distributed repeated measurements under some special assumptions about the form in which the data become available. Tang, Gnecco & Geller (1989) described a group sequential design for multivariate normal data in the absence of missing values, which uses previously generated tables for group sequential procedures with single endpoints. Wei, Su & Lachin (1990) showed how to conduct interim analyses with continuous or discrete repeated measurements using the regression method of Liang & Zeger (1986).

The present paper takes a nonparametric approach to the problem of repeated significance tests for detecting the stochastic ordering of two multivariate distributions. We are mainly concerned with multiple time-to-failure endpoints. The failures may be events of different natures or may be recurrences of the same kind of phenomena. The proposed method is also applicable to multiple quantitative measurements other than failure time variables. Arbitrary patterns of censorship and missing values are allowed.

The new proposal is based on the linear rank statistics with respect to the marginal distributions of the multiple endpoints. These marginal statistics are combined linearly to form a test statistic at a given look. The linear coefficients can be chosen to maximize asymptotic power against certain local alternatives. The critical values at various looks are obtained from the asymptotic joint distribution of the combined test statistics over time. An example taken from an AIDS clinical trial is presented. Several generalizations of the proposed method are also described.

2. CONSTRUCTION OF SEQUENTIAL TESTING PROCEDURE

Suppose that we wish to test the null hypothesis H_0 that the multivariate failure time distributions of treatments A and B are equal against a class of alternative hypotheses H_1 that the multivariate failure time vector of treatment A is stochastically larger or smaller than that of treatment B. For example, if we let $F_k^A(.)$ and $F_k^B(.)$ $(k=1,\ldots,K)$ be the marginal failure time distributions of treatments A and B, respectively, then H_1 may be that $F_k^A(.) \leqslant F_k^B(.)$ $(k=1,\ldots,K)$ with at least one strict inequality.

The notation to be introduced here is similar to that of Tsiatis (1982). We assume that n patients will enter the study. For $i=1,\ldots,n$ and $k=1,\ldots,K$, let Y_i denote the real time of entry for the ith patient, let V_{ki} denote the time from entry to the kth type of failure for the ith patient, and let C_{ki} denote the time from entry to censoring for the ith patient with respect to the kth failure time variable. Also, let Z_i be the indicator function of treatment A for the ith patient. The entry time variable Y_i, the failure time vector $V_i=(V_{1i},\ldots,V_{Ki})$ and the censoring time vector $C_i=(C_{1i},\ldots,C_{Ki})$ are assumed to be independent conditional on Z_i. In addition, (Y_i, V_i, C_i, Z_i) $(i=1,\ldots,n)$ are regarded as n independent and identically distributed random vectors.

When the data are reviewed at time t, for $i=1,\ldots,n$ and $k=1,\ldots,K$, we observe the time to failure or censoring $X_{ki}(t)$ and the failure indicator $\Delta_{ki}(t)$, where

$$X_{ki}(t)=\max\{\min(V_{ki}, t-Y_i, C_{ki}), 0\},$$

and $\Delta_{ki}(t)=1$ if $V_{ki} \leqslant \min(t-Y_i, C_{ki})$ and $\Delta_{ki}(t)=0$ otherwise. If V_{ki} is missing, we let both $X_{ki}(t)$ and $\Delta_{ki}(t)$ be zero.

At time t, the linear rank statistic with respect to the kth failure time variable can be written as

$$U_k(t)=\sum_{i=1}^{n}\Delta_{ki}(t)Q_k\{t, X_{ki}(t)\}\left[Z_i-\frac{S_k^{(1)}\{t, X_{ki}(t)\}}{S_k^{(0)}\{t, X_{ki}(t)\}}\right],$$

where

$$S_k^{(r)}(t, x) = n^{-1} \sum_j I\{X_{kj}(t) \geqslant x\} Z_j^r \quad (r = 0, 1),$$

and $I(.)$ is the indicator function. For fixed t, the random function $Q_k(t, x)$ is assumed to converge to a bounded function $q_k(t, x)$ in probability in sup norm over $[0, \tau]$ for every $\tau < t$. The statistic $U_k(t)$ corresponds to the log rank statistic (Mantel, 1966; Cox, 1972) if $Q_k(t, x) = 1$ and to the Peto-Prentice generalization of the Wilcoxon statistic (Peto & Peto, 1972; Prentice, 1978) if

$$Q_k(t, x) = \prod \{S_k^{(0)}\{t, X_{kj}(t)\}/[S_k^{(0)}\{t, X_{kj}(t)\} + n^{-1}]\}^{N_{kj}(t,x)},$$

where the product is over $j = 1, \ldots, n$, and where $N_{kj}(t, x) = I\{X_{kj}(t) \leqslant x, \Delta_{kj}(t) = 1\}$.

The marginal linear rank statistics $U_k(t)$ $(k = 1, \ldots, K)$ are generally correlated. An easy and natural way to combine these statistics for testing H_0 against H_1 is to construct a linear combination of the $U_k(t)$'s at each look. To perform sequential testing with such combined statistics, we need to derive the joint distribution of the statistics $U_k(t)$'s over failure time variables and over looks. Our approach is to approximate each of these statistics by a sum of n independent and identically distributed random variables so that the required asymptotic distribution can be obtained by the application of the multivariate central limit theorem. All subsequent results are derived under the null hypothesis unless specified otherwise.

Let

$$w_{ki}(t) = \int_0^t q_k(t, x)\left\{Z_i - \frac{s_k^{(1)}(t, x)}{s_k^{(0)}(t, x)}\right\} dM_{ki}(t, x),$$

where

$$s_k^{(r)}(t, x) = E\{S_k^{(r)}(t, x)\} \quad (r = 0, 1),$$

$$M_{ki}(t, x) = N_{ki}(t, x) - \int_0^x I\{X_{ki}(t) \geqslant u\} \lambda_k(u) du,$$

and $\lambda_k(.)$ is the common hazard function of the kth failure time variable. For fixed t, $M_{ki}(t, x)$ is a square integrable martingale with respect to the filtration

$$\mathcal{F}_{ki}(t) = \{\mathcal{F}_{ki}(t, x);\ 0 \leqslant x \leqslant t\},$$

where $\mathcal{F}_{ki}(t, x)$ is the sub σ-field generated by the random variables

$$\{I(Y_i \leqslant t), Z_i I(Y_i \leqslant t), Y_i I(Y_i \leqslant t), I(V_{ki} \leqslant \min(u, t - Y_i, C_{ki})),$$
$$I(C_{ki} \leqslant \min(u, t - Y_i, V_{ki})): 0 \leqslant u \leqslant x\}.$$

The random function $Q_k(t, x)$ is assumed to be a predictable process with respect to $\{\mathcal{F}_{ki}(t);\ i = 1, \ldots, n\}$.

Now, let $\tilde{U}_k(t) = \Sigma_i w_{ki}(t)$, which is a sum of n independent and identically distributed random variables. Application of martingale stochastic integral theorems shows that $n^{-\frac{1}{2}}\{U_k(t) - \tilde{U}_k(t)\}$ converges in probability to 0 and that $E\{w_{k1}(t)\} = 0$. These results are analogous to Lemmas 4 and 5 of Harrington, Fleming & Green (1982). It then follows from the multivariate central limit theorem and the Cramér-Wold device that

$$n^{-\frac{1}{2}}\{U_1(t_1), \ldots, U_K(t_1), \ldots, U_1(t_J), \ldots, U_K(t_J)\} \quad (t_1 < t_2 < \ldots < t_J)$$

converges to a multivariate normal distribution with mean 0 and with covariance element

$$\sigma_{kl}(t, t') = E\{w_{k1}(t) w_{l1}(t')\} \quad (k, l = 1, \ldots, K;\ t' \geqslant t).$$

As a referee pointed out, an independent increment structure in t arises, that is $\sigma_{kk}(t, t') = E\{w_{k1}(t)^2\}$ for $t' \geq t$, whenever the limiting weight function $q_k(t, x)$ is independent of t. Such is the case for the log rank and Peto-Prentice-Wilcoxon statistics and, more generally, for the G^ρ statistics of Harrington & Fleming (1982), but not for the Gehan-Wilcoxon statistic (Gehan, 1965).

It is natural to estimate $\sigma_{kl}(t, t')$ by $\hat{\sigma}_{kl}(t, t') = n^{-1} \Sigma_i W_{ki}(t) W_{li}(t')$, where

$$W_{ki}(t) = \Delta_{ki}(t) Q_k\{t, X_{ki}(t)\}\left[Z_i - \frac{S_k^{(1)}\{t, X_{ki}(t)\}}{S_k^{(0)}\{t, X_{ki}(t)\}}\right]$$
$$- \sum_{j=1}^{n} \frac{I\{X_{kj}(t) \leq X_{ki}(t)\}\Delta_{kj}(t) Q_k\{t, X_{kj}(t)\}}{nS_k^{(0)}\{t, X_{kj}(t)\}}\left[Z_i - \frac{S_k^{(1)}\{t, X_{kj}(t)\}}{S_k^{(0)}\{t, X_{kj}(t)\}}\right].$$

Note that $W_{ki}(t)$ is obtained from $w_{ki}(t)$ by substituting $q_k(t, x)$ by $Q_k(t, x)$, $s_k^{(r)}(t, x)$ by $S_k^{(r)}(t, x)$ $(r = 0, 1)$, and $\lambda_k(x)dx$ by $n^{-1}d\{\Sigma_i N_{ki}(t, x)\}/S_k^{(0)}(t, x)$. The consistency of $\hat{\sigma}_{kl}(t, t')$ can be established by the properties of empirical distribution functions and some probability arguments; see Wei & Lachin (1984) for a similar proof.

At time t, a weighted sum of the K standardized marginal linear rank statistics is expressed as $R(t) = \Sigma_k p_k(t)\{n^{-\frac{1}{2}} U_k(t)/\hat{\sigma}_{kk}^{\frac{1}{2}}(t, t)\}$. The weights $p_k(t)$ $(k = 1, \ldots, K)$ may be data-dependent and are assumed to converge in probability to deterministic quantities. Clearly, the covariance between $R(t)$ and $R(t')$ $(t' \geq t)$ can be consistently estimated by

$$\hat{\psi}(t, t') = \sum_k \sum_l p_k(t) p_l(t') \hat{\sigma}_{kl}(t, t')/\{\hat{\sigma}_{kk}(t, t)\hat{\sigma}_{ll}(t', t')\}^{\frac{1}{2}}.$$

In addition, the test statistic $R(t)/\hat{\psi}^{\frac{1}{2}}(t, t)$ is asymptotically standard normal.

The optimal choice of the weights $p(t) = (p_1(t), \ldots, p_K(t))$ depends on the alternative hypothesis anticipated in the experiment. For an illustration, we consider a sequence of local alternative hypotheses

$$H_{1n}: \lambda_{kn}^A(x) = \lambda_k^B(x) \exp\{n^{-\frac{1}{2}}\beta g_k(x)\} \quad (k = 1, \ldots, K),$$

where $\lambda_{kn}^A(.)$ is the kth hazard function for treatment A with sample size n, $\lambda_k^B(.)$ is the kth hazard function for treatment B, $g_k(.)$ is a known function of time, and β is an arbitrary nonzero constant. Suppose that the random function $Q_k(t, x)$ is chosen such that $q_k(t, x) = g_k(x)$ for fixed t, which is optimal for the marginal test statistic (Gill, 1980, Ch. 5; Schoenfeld, 1981). Then, under H_{1n}, $n^{-\frac{1}{2}} U_k(t)/\hat{\sigma}_{kk}^{\frac{1}{2}}(t, t)$ is asymptotically normal with mean $\beta\sigma_{kk}^{\frac{1}{2}}(t, t)$ and variance 1. In addition, the test statistic $R(t)/\hat{\psi}^{\frac{1}{2}}(t, t)$ is asymptotically normal with unit variance and with a mean which is asymptotically equivalent to $\beta \Sigma_k p_k(t)\hat{\sigma}_{kk}^{\frac{1}{2}}(t, t)/\hat{\psi}^{\frac{1}{2}}(t, t)$. Hence, the asymptotic power of the test statistic $R(t)/\hat{\psi}^{\frac{1}{2}}(t, t)$ against H_{1n} is maximized if $p(t) = \hat{\Lambda}(t)^{-1}\hat{\eta}(t)$, where

$$\hat{\eta}(t) = (\hat{\sigma}_{11}^{\frac{1}{2}}(t, t), \ldots, \hat{\sigma}_{KK}^{\frac{1}{2}}(t, t))', \quad \hat{\Lambda}(t) = \left\{\frac{\hat{\sigma}_{kl}(t, t)}{\{\hat{\sigma}_{kk}(t, t)\hat{\sigma}_{ll}(t, t)\}^{\frac{1}{2}}}; k, l = 1, \ldots, K\right\}$$

(Rao, 1973, p. 60).

The relative weighting of the marginal test statistics through the aforementioned calculation of the vector $p(t)$ is heavily driven by the information available for each statistic, without formal consideration given to the relative clinical importance of the outcome measures. One simple method of attaching unequal priorities to various endpoints is to incorporate different weights into $\hat{\eta}(t)$. However, one should specify such weights on an *a priori* basis in order to avoid bias.

Suppose that repeated tests based on $R(t)$ are to be performed at time points $t_1 < \ldots < t_J$ with an overall significance level α. Then we reject H_0 at t_j if the observed absolute value of the test statistic $T_j = R(t_j)/\hat{\psi}^{\frac{1}{2}}(t_j, t_j)$ exceeds d_j. The boundary values d_j $(j = 1, \ldots, J)$ are obtained by the method of Slud & Wei (1982). Specifically, these values are determined recursively by the following equations:

$$\text{pr}\{|G_1| < d_1, \ldots, |G_{j-1}| < d_{j-1}, |G_j| > d_j\} = \alpha_j \quad (j = 1, \ldots, J),$$

where $\{\alpha_1, \ldots, \alpha_J\}$ is a sequence of exit probabilities such that $\Sigma\, \alpha_j = \alpha$, and $(G_1, \ldots, G_j)$ is a zero-mean multivariate normal with covariance matrix

$$\{\hat{\psi}(t_a, t_b)/\{\hat{\psi}(t_a, t_a)\hat{\psi}(t_b, t_b)\}^{\frac{1}{2}};\ a, b = 1, \ldots, j\}.$$

The choice of $\{\alpha_1, \ldots, \alpha_J\}$ is as intrinsically related to the power of test procedures in a group sequential design as are the limiting weight functions $\{q_1(.,.), \ldots, q_K(.,.)\}$. A more complete consideration of the asymptotic power against certain local alternatives would involve the α_j's. We generally recommend that an increasing sequence of α_j be used so that the resulting sequential procedure would not only allow early detection of substantial treatment differences but also have power close to that of the corresponding single stage procedure.

3. An example

This research was motivated by AIDS clinical trials evaluating therapies in the treatment of patients with human immunodeficiency virus. The major clinical events of interest in AIDS trials are opportunistic infections and deaths. Opportunistic infections are potentially life-threatening infections that occur when the immune system is compromised. In a typical AIDS trial, opportunistic infections are more frequently observed than deaths.

A double-blind, placebo-controlled clinical trial on the efficacy of oral azidothymidine, AZT, for treating patients with AIDS or AIDS-related complex was conducted in 1986 (Fischl et al., 1987). The trial was terminated prematurely due to early evidence of substantial treatment difference in mortality. In many other AIDS clinical trials, group sequential tests have been performed on the time to the first major clinical event, which make little use of the mortality information because virtually all deaths are preceded by opportunistic infections. Here, we show how the method developed in the last section could have been applied to the AZT trial by regarding the time to the first infection and the time to death as a bivariate failure time vector.

Two hundred and eighty-one patients were enrolled in the AZT trial between February and June 1986, among whom 144 were assigned to AZT and 137 to placebo. The study was terminated in September 1986. Forty-six subjects were withdrawn from the trial before its termination. The main reasons for withdrawal were the occurrence of opportunistic infections and the patient's request. Subsequent information about these patients was lost, but their death times were always recorded. For this reason, the individual withdrawal time is treated as a censoring variable for the first opportunistic infection but not for death in our analysis. The small number of subjects who had been withdrawn from the study before developing opportunistic infections might provide informative censoring with respect to the infection time. All the deaths reported in this study occurred after the first infections or withdrawals. The data from the AZT trial are summarized in Table 1.

Suppose that the investigators had decided to examine the data of the first infections and deaths at $t_1 = 3$ months, $t_2 = 5$ months, $t_3 = 7$ months, $t_4 = 9$ months and $t_5 = 11$ months

Table 1. *Summary of the* AZT *trial data*

		Time intervals (months) 0-3	3-5	5-7
AZT group	Entrants	88	56	0
	First infections	7	12	6
	Deaths	0	0	1
Placebo group	Entrants	88	49	0
	First infections	10	19	22
	Deaths	3	4	11

of the study by using combined log rank tests with exit probabilities $\alpha_1 = 0{\cdot}005$, $\alpha_2 = 0{\cdot}005$, $\alpha_3 = 0{\cdot}01$, $\alpha_4 = 0{\cdot}01$ and $\alpha_5 = 0{\cdot}02$. The observed log rank statistics $U_k(t)$ $(k = 1, 2)$ at $t = t_1$, t_2 and t_3 along with the their estimated variance-covariance matrix are displayed in Table 2. The corresponding standardized statistics are shown in the first two rows of Table 3.

For illustrations, we consider two different weighting schemes: (i) $p(t) = \hat{\Lambda}(t)^{-1}\hat{\eta}(t)$, which is asymptotically optimal against H_{1n} if $g_1(x) = g_2(x) = 1$; and (ii) $p_1(t) = p_2(t)$. With scheme (i), the first infections and deaths are weighted by about 4 to 1 at t_1, t_2 and t_3, the main reason for the domination of the first infections being that there were far more observed infections than observed deaths. The values of the test statistics T_j $(j = 1, \ldots, 3)$ for both weighting schemes are shown in Table 3. Scheme (ii) provides stronger evidence for the benefit of AZT than scheme (i) and separate log rank statistics. The approximate correlation matrices

$$\{\hat{\psi}(t_a, t_b)/\{\hat{\psi}(t_a, t_a)\hat{\psi}(t_b, t_b)\}^{\frac{1}{2}};\ a, b = 1, \ldots, 3\}$$

for weighting schemes (i) and (ii) are, respectively,

$$\begin{bmatrix} 1 & . & . \\ 0{\cdot}608 & 1 & . \\ 0{\cdot}444 & 0{\cdot}764 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & . & . \\ 0{\cdot}664 & 1 & . \\ 0{\cdot}443 & 0{\cdot}704 & 1 \end{bmatrix}.$$

Table 2. *Marginal log rank statistics and estimated variance-covariance matrix*

			$n\hat{\sigma}_{kl}(t, t')$ $t' = 3$		$t' = 5$		$t' = 7$	
t	k	$U_k(t)$	$l = 1$	$l = 2$	$l = 1$	$l = 2$	$l = 1$	$l = 2$
3	1	−1·365	4·160	0·674	4·194	0·625	4·045	0·672
3	2	−1·474		0·714	0·725	0·736	0·701	0·754
5	1	−6·021			12·051	0·857	11·859	1·690
5	2	−3·631				1·866	0·820	1·852
7	1	−16·001					19·561	2·737
7	2	−8·920						5·061

Table 3. *Standardized log rank test statistics calculated over time*

	$t = 3$	$t = 5$	$t = 7$
First infections alone	−0·669	−1·734	−3·618
Deaths alone	−1·745	−2·658	−3·965
First infections and deaths			
(i) $p(t) = \hat{\Lambda}(t)^{-1}\hat{\eta}(t)$	−0·709	−2·199	−4·245
(ii) $p_1(t) = p_2(t)$	−1·447	−2·858	−4·748

By the normal integration algorithm of Schervish (1984), we obtain {2·807, 2·765, 2·496} and {2·807, 2·753, 2·510} as the sequences of critical values $\{d_1, d_2, d_3\}$ for weighting schemes (i) and (ii), respectively. Therefore, the trial would have been terminated at t_2, two months earlier than the actual termination date, had the hypothetical stopping rule with weighting scheme (ii) been enforced. In contrast, for the same sequence of exit probabilities, sequential testing based on weighting scheme (i), using the first infections alone or deaths alone, would not have been significant until t_3.

4. Remarks

We have presented a simple sequential testing procedure for clinical trials designed with multiple endpoints under some realistic assumptions about the form in which the data become available. The proposed method is useful when the endpoints being combined are biologically related and demonstrate treatment differences in the same direction. A FORTRAN program which implements the new procedure is available from the author.

The assumption of noninformative censoring may be violated if the occurrence of one endpoint precludes the development of others. One possible solution to this problem is to redefine the endpoints. In the AZT trial, for instance, it would have been more appropriate to define the double endpoints as the time to the first major event and the time to death had there been any first infections being censored by deaths.

Covariate Z has been defined as a treatment indicator. The proposed method is still applicable if Z is continuous. In addition, the extension to the case of a time-varying covariate is straightforward.

Stratified linear rank statistics have been demonstrated to have adequate power for comparing survival data from two treatment groups in heterogeneous populations (Lininger et al. 1979). A sequential testing procedure based on the weighted sum of the marginal stratified linear rank statistics can be easily developed from the basic ideas presented in § 2.

An alternative way of accounting for the heterogeneity is to model each marginal distribution of the multivariate failure time vector by a Cox proportional hazards model (Wei, Lin & Weissfeld, 1989) and then calculate the marginal partial likelihood score statistics for testing no treatment effects in the presence of other covariates. To construct a sequential testing procedure with the linear combination of these marginal score statistics, we again approximate each one of them by a sum of n independent and identically distributed random variables. This can be accomplished by the techniques used by Lin & Wei (1989) in deriving the robust score test for the Cox regression model. The resulting procedure should be insensitive to model misspecification.

Repeated confidence intervals for a single parameter of interest such as the common hazard ratio in the marginal proportional hazards models or the common scale-change parameter in the marginal accelerated failure time models can be constructed by inverting the proposed sequential tests with some simple modifications on the marginal statistics. The readers are referred to Jennison & Turnbull (1989, § 4.1.2) and Lin & Wei (1991) for similar operations with univariate failure time data.

An important application of the new method is to the problem of multiple endpoints other than failure time variables with possible missing values. These multivariate observations may be measurements of several characteristics or may be repeated measurements of the same characteristic at different occasions. For $k = 1, \ldots, K$ and $i = 1, \ldots, n$, let $X_{ki}(t) = V_{ki}$ and $\Delta_{ki}(t) = 1$ if the measurement of interest V_{ki} is available at time t, and

let $X_{ki}(t) = -\infty$ and $\Delta_{ki}(t) = 0$ otherwise. Then the method developed in § 2 can be directly applied to the current setting provided that the process that generates missing data is ignorable (Rubin, 1976).

ACKNOWLEDGEMENTS

The author is deeply grateful to the Burroughs Wellcome Company for providing the AZT trial data and to a referee for many insightful comments and constructive suggestions. He would also like to thank Drs David Schoenfeld, Michael Kosorok, L. J. Wei, John Andrews and Michael Wulfsohn for their helpful discussions. This research was supported by National Institutes of Health grants. The original manuscript was completed while the author was at Harvard University.

REFERENCES

COX, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B **34**, 187-220.

FISCHL, M. A. ET AL. (1987). The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *N. Eng. J. Med.* **317**, 185-91.

GEARY, D. N. (1988). Sequential testing in clinical trials with repeated measurements. *Biometrika* **75**, 311-8.

GEHAN, E. A. (1965). A generalized two-sample Wilcoxon test for comparing arbitrarily singly-censored samples. *Bometrika* **52**, 203-23.

GILL, R. D. (1980). *Censoring and Stochastic Integrals.* Amsterdam: The Mathematical Centre.

HARRINGTON, D. P. & FLEMING, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553-66.

HARRINGTON, D. P., FLEMING, T. R. & GREEN, S. J. (1982). Procedures for serially testing general hypothesis in censored survival data. In *Survival Analysis*, Ed. J. Crowley and R. Johnson, pp. 269-86. Hayward: Institute of Mathematical Statistics.

JENNISQN, C. & TURNBULL, B. W. (1989). Interim analyses: the repeated confidence interval approach (with discussion). *J. R. Statist. Soc.* B **51**, 305-61.

LIANG, K. Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

LIN, D. Y. & WEI, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Am. Statist. Assoc.* **84**, 1074-8.

LIN, D. Y. & WEI, L. J. (1991). Repeated confidence intervals for a scale change in a sequential survival study. *Biometrics.* To appear.

LININGER, L., GAIL, M. H., GREEN, S. B. & BYAR, D. P. (1979). Comparison of four tests for equality of survival curves in the presence of stratification and censoring. *Biometrika* **66**, 419-28.

MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemo. Rep.* **50**, 163-70.

O'BRIEN, P. C. (1984). Procedures for combining samples with multiple endpoints. *Biometrics* **40**, 1079-87.

PETO, R. & PETO, J. (1972). Asymptotically efficient rank invariance test procedures (with discussion). *J. R. Statist. Soc.* A **135**, 185-206.

POCOCK, S. J., GELLER, N. L. & TSIATIS, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487-98.

PRENTICE, R. L. (1978). Linear rank tests with censored data. *Biometrika* **65**, 167-79.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.

RUBIN, D. (1976). Inference and missing data. *Biometrika* **63**, 581-92.

SCHERVISH, M. J. (1984). Multivariate normal probabilities with error bound. *Appl. Statist.* **33**, 81-94. Corrections (1985), **34**, 103-4.

SCHOENFELD, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68**, 316-9.

SLUD, E. & WEI, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Am. Statist. Assoc.* **77**, 862-8.

TANG, D.-I., GNECCO, C. & GELLER, N. L. (1989). Design of group sequential clinical trials with multiple endpoints. *J. Am. Statist. Assoc.* **84**, 776-9.

TSIATIS, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *J. Am. Statist. Assoc.* **77**, 855-61.

WEI, L. J. & JOHNSON, W. E. (1985). Combining dependent tests with incomplete repeated measurements. *Biometrika* **72**, 359–64.

WEI, L. J. & LACHIN, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *J. Am. Statist. Assoc.* **79**, 653–61.

WEI, L. J., LIN, D. Y. & WEISSFELD, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Statist. Assoc.* **84**, 1065–73.

WEI, L. J., SU, J. Q. & LACHIN, J. M. (1990). Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika* **77**, 359–64.

[*Received February* 1990. *Revised July* 1990]

NOTE ADDED IN PROOF

Since this paper went to press, an unpublished manuscript by J. Q. Su and J. M. Lachin has been brought to my attention. This manuscript describes group sequential methods based on the multivariate U-statistics of Wei & Johnson (1985).