Repeated Confidence Intervals for a Scale Change in a Sequential Survival Study

D. Y. Lin; L. J. Wei

Biometrics, Vol. 47, No. 1. (Mar., 1991), pp. 289-294.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28199103%2947%3A1%3C289%3ARCIFAS%3E2.0.CO%3B2-8

Biometrics is currently published by International Biometric Society.

BIOMETRICS 47, 289–294
March 1991

Repeated Confidence Intervals for a Scale Change in a Sequential Survival Study

D. Y. Lin

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

and

L. J. Wei

Department of Statistics, University of Wisconsin, 1210 W. Dayton Street, Madison, Wisconsin 53706, U.S.A.

SUMMARY

This note considers the problem of comparing the survival distributions of two treatment groups in a clinical trial where the treatment difference is measured in terms of a time-scale change. Patients enter the study sequentially and may be subject to random loss to follow-up. The survival data are reviewed periodically for early evidence of the treatment difference. An approach to constructing repeated confidence intervals for the scale-change parameter is described. These intervals provide useful information about the magnitude of the treatment difference in interim analyses. An example taken from an AIDS clinical trial is given.

1. Introduction

In a clinical trial comparing the survival distributions of two treatment groups, suppose that patients enter the study serially and that the failure time observations, which may be censored, are reviewed periodically. A few sequential testing procedures are available for monitoring such a long-term survival study (e.g., Jones and Whitehead, 1979; Slud and Wei, 1982; Tsiatis, 1982). In some situations, however, significance tests do not provide enough information for reaching a decision on the early termination of a trial. Periodic evaluations of the magnitude of the treatment difference can be very useful.

Jennison and Turnbull (1984) and Lai (1984) have advocated the use of repeated confidence intervals (RCIs), which are obtained by inverting group sequential tests. This approach allows a full exploration of the data at each interim analysis and does not depend on a rigidly enforced statistical stopping rule. The final member of the sequence of RCIs may also be used to summarize the information about the parameter of interest in the final report upon termination of a sequential trial. In the context of comparing two survival distributions, Jennison and Turnbull (1989) described methods of constructing RCIs for the hazard ratio under the proportional hazards assumption (Cox, 1972).

A useful alternative to the proportional hazards model is the accelerated failure time model, which relates covariates linearly to the logarithm of the survival time (Kalbfleisch and Prentice, 1980, pp. 32–34). These two models are identical when the underlying distribution is Weibull. The accelerated failure time modeling is especially appealing to medical investigators due to its straightforward interpretation. Under this model, the treatment difference in the two-sample case is measured in terms of a time-scale change, say θ_0. Specifically, if V_1 and V_2 are the

Key words: Accelerated failure time model; Clinical trial; Group sequential design; Interim analyses; Linear rank statistic; Repeated significance tests.

underlying failure time variables for treatments A and B, respectively, then V_2/θ_0 has the same distribution as V_1. It may be desirable to construct repeated confidence intervals for θ_0 in a sequential survival study.

Single-stage nonparametric estimation of the scale-change parameter θ_0 with censored observations has been studied by Louis (1981) and Wei and Gail (1983). Their ideas are as follows. Suppose that random samples of survival times $(V_{11}, \ldots, V_{1n_1})$ and $(V_{21}, \ldots, V_{2n_2})$ come from distributions $F_1(v) = F(v)$ and $F_2(v) = F(v/\theta_0)$, respectively, where F and θ_0 are unknown. For $i = 1, \ldots, n_g$ and $g = 1, 2$, one observes X_{gi} and Δ_{gi}, where $X_{gi} = \min(V_{gi}, W_{gi})$, and $\Delta_{gi} = 1$ if $X_{gi} = V_{gi}$ and $\Delta_{gi} = 0$ otherwise. The censoring variables W_{gi} are mutually independent and are also independent of V_{gi}. Let $U(\mathbf{X}_1, \mathbf{X}_2, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$ be a two-sample statistic for testing H_0: $\theta_0 = 1$, where $\mathbf{X}_g = (X_{g1}, \ldots, X_{gn_g})$ and $\boldsymbol{\Delta}_g = (\Delta_{g1}, \ldots, \Delta_{gn_g})$ for $g = 1, 2$. Suppose that the null distribution of this statistic is asymptotically normal with zero mean. Then, for any θ_0, the statistic $U(\mathbf{X}_1, \mathbf{X}_2/\theta_0, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$ is also asymptotically normal with zero mean. This fact motivates a test-based estimator $\hat{\theta}$ as the value of θ for which $U(\mathbf{X}_1, \mathbf{X}_2/\theta, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$ is as close to zero as possible. In addition, test-based confidence intervals for θ_0 can be obtained from the large-sample distribution of $U(\mathbf{X}_1, \mathbf{X}_2/\theta_0, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$.

Based on the aforementioned ideas, repeated confidence intervals for the scale-change parameter θ_0 are derived in the next section. An example taken from an AIDS clinical trial is given in Section 3.

2. Construction of Repeated Confidence Intervals

Let the nonnegative random variable Y denote the real time of entry, let the continuous nonnegative random variable V denote the time to failure measured from the entry time, and let the nonnegative random variable W denote the time to censoring, also measured from the entry time. In addition, let Z be the indicator variable of treatment A. The random variables Y, V, and W are assumed to be independent given Z.

When the data are examined at time t, we observe the time to failure or censoring $X(t)$ and the failure indicator $\Delta(t)$, where $X(t) = \max\{\min(V, t - Y, W), 0\}$, and $\Delta(t) = 1$ if $V \leqslant \min(t - Y, W)$ and $\Delta(t) = 0$ otherwise. At time t, the data consist of n independent and identically distributed replicates of $\{X(t), \Delta(t), Z\}$, where n is the total number of patients *ever* entering the study. It will become evident in the sequel that the statistics calculated at time t depend only on the data of the patients who have entered the study by that time point. For convenience, we index these triplets $\{X(t), \Delta(t), Z\}$ according to the orders of patients' entry times.

Now, assume that the survival distributions of treatments A and B are F_1 and F_2, respectively. Let us rescale the possibly censored failure times from treatment B by dividing them by θ, and let $\tilde{X}_i(t, \theta)$ denote $X_i(t)/\theta^{1-Z_i}$ for $i = 1, \ldots, n(t)$, where $n(t)$ is the number of patients already in the study by time t. We consider the following class of two-sample rank statistics:

$$U(t, \theta) = \sum_{i=1}^{n(t)} \Delta_i(t) Q\big[t, \tilde{X}_i(t, \theta)\big]\big\{Z_i - \bar{Z}\big[t, \tilde{X}_i(t, \theta)\big]\big\},$$

where $\bar{Z}(t, x) = \sum_{j=1}^{n(t)} I[\tilde{X}_j(t, \theta) \geqslant x] Z_j / N(t, x)$, $N(t, x) = \sum_{j=1}^{n(t)} I[\tilde{X}_j(t, \theta) \geqslant x]$, and $I(\cdot)$ is the indicator function. The random function $Q(t, x)$ corresponds to the weighting function of the test statistics studied by Tarone and Ware (1977) and Prentice and Marek (1979). For example, $Q(t, x) = 1$ for the log-rank statistic, and

$$Q(t, x) = \prod_{j=1}^{n(t)} \big\{N\big(t, \tilde{X}_j(t, \theta)\big)/\big[N\big(t, \tilde{X}_j(t, \theta)\big) + 1\big]\big\}^{I[\tilde{X}_j(t, \theta) \leqslant x,\ \Delta_j(t) = 1]}$$

for the generalized Wilcoxon statistic proposed by Peto and Peto (1972) and Prentice (1978).

It follows from Corollary 1 of Wei and Gail (1983) that a consistent estimator of θ_0 at time t is given by

$$\hat{\theta}(t) = \left[\sup\{\theta: U(t,\theta) > 0\} + \inf\{\theta: U(t,\theta) < 0\}\right]/2.$$

The above function is well defined because $U(t,\theta)$ is a nonincreasing function of θ for any fixed t.

Suppose that the data are reviewed at time points $t_1 < \cdots < t_K$. If there is a moderate number of events by t_1, the joint distribution of the random vector $\{U(t_1,\theta_0), \ldots, U(t_K,\theta_0)\}$ can be approximated by a zero-mean multivariate normal distribution. In addition, a typical element $\sigma_{kl}(\theta_0)$ $(k \leqslant l)$ of the corresponding covariance matrix is

$$\sum_{i=1}^{n(t_k)} \Delta_i(t_k) Q\left[t_k, \tilde{X}_i(t_k,\theta_0)\right] Q\left[t_l, \tilde{X}_i(t_k,\theta_0)\right]$$

$$\times \sum_{j=1}^{n(t_k)} I\left[\tilde{X}_j(t_k,\theta_0) \geqslant \tilde{X}_i(t_k,\theta_0)\right]\left\{Z_j - \bar{Z}\left[t_k, \tilde{X}_i(t_k,\theta_0)\right]\right\}^2 / N\left[t_k, \tilde{X}_i(t_k,\theta_0)\right].$$

The forgoing results follow from Theorems 3.1 and 3.2 of Tsiatis (1982). Under rather mild conditions (Wei and Gail, 1983), $n^{-1}\sigma_{kl}(\theta_0)$ is asymptotically equivalent to $n^{-1}\sigma_{kl}(\hat{\theta})$ for a consistent estimator $\hat{\theta}$.

It is easy to show that the process $U(t,\theta_0)$ has asymptotically independent increments if the log-rank or the Peto–Prentice–Wilcoxon weighting function is used. However, this property does not hold for Gehan's statistic unless all patients enter the study simultaneously.

Asymptotically distribution-free repeated confidence intervals for θ_0 with $(1-\alpha)$ simultaneous coverage probability can be obtained by inverting a sequence of tests based on $\{U(t_k,\theta); k = 1, \ldots, K\}$ for testing $\theta = \theta_0$ with an overall Type I error probability α. In particular, the kth confidence interval $(\underline{\theta}_k, \bar{\theta}_k)$ $(k = 1, \ldots, K)$ has limits

$$\underline{\theta}_k = \inf\{\theta: \sigma_{kk}^{-1/2}\left[\hat{\theta}(t_k)\right] U(t_k,\theta) \leqslant c_k\}$$

and

$$\bar{\theta}_k = \sup\{\theta: \sigma_{kk}^{-1/2}\left[\hat{\theta}(t_k)\right] U(t_k,\theta) \geqslant -c_k\}.$$

The boundary values c_k are constructed recursively with a sequence of prespecified exit probabilities $\pi_1, \ldots, \pi_K$, where $\sum_{k=1}^{K}\pi_k = \alpha$ (Slud and Wei, 1982). Specifically, if $c_1, \ldots, c_{k-1}$ have been obtained, then c_k is determined by

$$\Pr\{|G_1[\hat{\theta}(t_k)]| < c_1, \ldots, |G_{k-1}[\hat{\theta}(t_k)]| < c_{k-1}, |G_k[\hat{\theta}(t_k)]| \geqslant c_k\} = \pi_k,$$

where $\{G_1(\theta), \ldots, G_k(\theta)\}$ is a zero-mean multivariate normal with covariance matrix $\{\sigma_{ab}(\theta)/[\sigma_{aa}(\theta)\sigma_{bb}(\theta)]^{1/2}; a, b = 1, \ldots, k\}$. An increasing sequence of π_k's is usually recommended so that the Kth repeated confidence interval is not drastically different from the customary fixed-sample-size confidence interval. The readers are referred to Jennison and Turnbull (1989, §3.3) for a fine discussion on the choice of π_k's.

The optimal choice of the weighting function $Q(t, x)$ depends on the unknown distribution function $F(x)$. It can be shown that the asymptotic variance of the estimator $\hat{\theta}(t)$ will be minimized if the probability limit of $Q(t, x)$ is proportional to $xd\{\log \lambda(x)\}/dx$, where $\lambda(x)$ is the hazard function of $F(x)$ (Tsiatis, 1990). (Thus, a constant weighting function should be used if Weibull distributions are anticipated.) This result is similar to that of finding the optimal weighting function for the linear rank statistic in hypothesis testing. Research is currently undertaken by Tsiatis and his colleagues to adaptively estimate $d\{\log \lambda(x)\}/dx$.

3. An Illustration

A double-blind, placebo-controlled trial on the efficacy of oral azidothymidine (AZT) for treating AIDS patients was conducted in 1986 (Fischl et al., 1987). Two hundred eighty-one subjects were enrolled in the trial between February and June 1986, among whom 144 were assigned to AZT and 137 to placebo. The study was terminated in September 1986. By the end of the study, 25 patients in the AZT group and 51 patients in the placebo group had developed at least one opportunistic infection. The opportunistic infection is a potentially life-threatening illness that occurs when a patient's immune system is compromised. The time from entering the study to the first opportunistic infection has been commonly used as the primary response variable in AIDS clinical trials. The history of patient enrollment and first opportunistic infections for the AZT study is summarized in Table 1. We now illustrate how the procedures developed in the last section could have been used to monitor this trial.

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
Placebo group	Entrants	88	49	0
	First infections	10	19	22

Suppose that the investigators had planned to review the data of the first infections at $t_1 = 3$ (months), $t_2 = 5$ (months), and $t_3 = 7$ (months) in the study. In addition, it had been decided that RCIs for the scale-change parameter θ_0 with $\pi_1 = .01$, $\pi_2 = .015$, and $\pi_3 = .025$ would be constructed using the log-rank statistics.

The point estimates $\hat{\theta}(t_k)$ at t_1, t_2, and t_3 are, respectively, .645, .653, and .519, which indicate that on average the time to the first infection after entering the trial for the AZT group was nearly twice as long as that for the placebo group. The approximate correlation matrices $\{\sigma_{ab}[\hat{\theta}(t_k)]/(\sigma_{aa}[\hat{\theta}(t_k)]\sigma_{bb}[\hat{\theta}(t_k)])^{1/2};\ a, b = 1, \ldots, k\}$ $(k = 2, 3)$ at t_2 and t_3 are, respectively,

$$\begin{bmatrix} 1 & \\ .6129 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & & \\ .6206 & 1 & \\ .5104 & .8224 & 1 \end{bmatrix}.$$

By applying the integration algorithm of Schervish (1984) to the zero-mean normal distributions with the above covariance matrices, we obtain $(2.576, 2.381, 2.097)$ as the sequence of boundary values $\{c_k;\ k = 1, 2, 3\}$. Then the RCIs for θ_0 are $(.000, 6.600)$, $(.168, 1.207)$, and $(.206, .733)$ at t_1, t_2, and t_3, respectively. The first RCI is extremely wide due to a large boundary value and a small number of events. The upper limit of the RCI at t_3 is far below the null value 1, which provides strong evidence for the benefit of AZT for the AIDS patients.

If the generalized Wilcoxon statistics are used instead of the log-rank statistics, then the point estimates for θ_0 at t_1, t_2, and t_3 are, respectively, .676, .674, and .530, and the corresponding RCIs are $(.000, 6.600)$, $(.183, 1.286)$, and $(.225, .757)$. These results are similar to those based on the log-rank statistics.

For comparison, we also construct the RCIs for the hazard ratio of AZT vs placebo by replacing the statistics $U(t, \theta)$ in Section 2 with the partial-likelihood score statistics under the Cox model. The confidence intervals for this ratio with the aforementioned exit probabilities π_k $(k = 1, 2, 3)$ are $(.121, 2.655)$, $(.270, 1.189)$, and $(.233, .677)$ at t_1, t_2, and t_3, respectively.

The Gill–Schumacher test of the proportional hazards assumption (Gill and Schumacher, 1987) with the log-rank and Peto–Prentice–Wilcoxon weighting functions based on the entire data set yields a P-value of .035. The corresponding P-value from a similar test for the scale-change model proposed by Wei, Ying, and Lin (1990) is .198. Thus, the scale-change model seems more appropriate than the proportional hazards model for this study.

4. Remarks

The purpose of this note was to provide repeated confidence intervals for the scale-change parameter θ_0 with a prescribed overall coverage probability. These intervals are useful not only in trial monitoring but also in the final report upon termination of the study. Several authors have derived confidence intervals for the parameters of the normal and binomial distributions following sequential testing. Such procedures are applicable only if an appropriate stopping rule is strictly enforced, and their use in the survival setting has yet to be explored. Reporting the current RCI on termination gives a somewhat conservative interval but allows greater flexibility as this interval is valid whatever stopping criterion is used.

Another estimation issue arising in a sequential survival study, i.e., the point estimation for θ_0 following repeated significance tests, was not addressed here. It is widely recognized that the maximum likelihood estimator calculated after sequential testing tends to overestimate the magnitude of treatment differences as a result of optional stopping, especially when sampling is terminated early (Cox, 1952). Kim (1988) and Hughes and Pocock (1988) proposed methods to reduce such bias for the cases of normal and binary responses. It would be valuable to carry out a similar investigation for the scale-change model.

ACKNOWLEDGEMENTS

This research was supported by the National Institutes of Health. The authors are grateful to the Shorter Communications editor, an associate editor, and referees for their useful comments.

RÉSUMÉ

Cette note traite du problème de la comparaison des distributions de survie entre deux groupes de traitement dans un essai clinique où la différence entre traitements est mesurée en terme de variation d'un paramètre d'échelle. Les patients rentrent dans l'étude séquentiellement; celle-ci peut comporter des perdus de vue. Les données de survie sont analysées périodiquement afin de mettre en évidence le plus tôt possible des différences entre traitements. Une méthode de construction d'intervalles de confiance répétés du paramètre d'échelle est décrite. Ces intervalles fournissent une information utile sur la grandeur de la différence entre traitements dans les analyses intérimaires. Un exemple est donné à partir d'un essai clinique sur le SIDA.

REFERENCES

Cox, D. R. (1952). A note on the sequential estimation of means. *Proceedings of the Cambridge Philosophical Society* **48**, 447–450.

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

Fischl, M. A., Richman, D. D., Grieco, M. H., et al. (1987). The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *New England Journal of Medicine* **317**, 185–191.

Gill, R. D. and Schumacher, M. (1987). A simple test of the proportional hazards assumption. *Biometrika* **74**, 289–300.

Hughes, M. D. and Pocock, S. J. (1988). Stopping rules and estimation problems in clinical trials. *Statistics in Medicine* **7**, 1231–1242.

Jennison, C. and Turnbull, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials* **5**, 33–45.

Jennison, C. and Turnbull, B. W. (1989). Interim analyses: The repeated confidence interval approach (with Discussion). *Journal of the Royal Statistical Society, Series B* **51**, 305–361.

Jones, D. and Whitehead, J. (1979). Sequential forms of the log rank and modified Wilcoxon tests for censored data. *Biometrika* **66**, 105–113.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kim, K. (1988). Improved approximation for estimation following closed sequential tests. *Biometrika* **75**, 121–128.

Lai, T. L. (1984). Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach. *Communications in Statistics, Series A* **13**, 2355–2368.

Louis, T. A. (1981). Nonparametric analysis of an accelerated failure time model. *Biometrika* **68**, 381–390.

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

Prentice, R. L. (1978). Linear rank tests with 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.

Schervish, M. J. (1984). Multivariate normal probabilities with error bound. *Applied Statistics* **33**, 81–94. Corrections, **34**, 103–104.

Slud, E. and Wei, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association* **77**, 862–868.

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

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* **77**, 855–861.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics* **18**, 354–372.

Wei, L. J. and Gail, M. H. (1983). Nonparametric estimation for a scale-change with censored observations. *Journal of the American Statistical Association* **78**, 382–388.

Wei, L. J., Ying, Z., and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**, 845–851.

Received September 1989; *revised June and September* 1990; *accepted October* 1990.