The Korean Communications
in Statistics Vol. 5, No. 2, 1998
pp. 459-475

A Study on the Group Sequential Methods for Comparing Survival Distributions in Clinical Trials[1)]

Jae Won Lee[2)]

Abstract

In many clinical trials, we are interested in comparing the failure time distribution of different treatment groups. Because of ethical and economic reasons, clinical trials need to be monitored for early dramatic benefits or potential harmful effects. Prior knowledge, evolving knowledge, statistical considerations, medical judgment and ethical principles are all involved in the decision to terminate a trial early, and thus the monitoring is usually carried out by an independent scientific committee. This paper reviews the recently proposed group sequential testing procedures for clinical trials with survival data. Design considerations of such clinical trials are also discussed. This paper compares the characteristics of each of these methods and provides the biostatisticians with the guidelines for choosing the appropriate group sequential methods in a given situation.

1. Introduction

While it would be much simpler statistically to carry out a clinical trial with just a single planned analysis of the data at an appropriately prespecified time, it is rarely done in major clinical trials. The investigator's responsibility to the study subjects demands that the results be monitored during the trial. If data indicate that the new treatment is harmful to the subjects, the investigators have obligation to terminate the trial early or modify the study in some acceptable way. If these data demonstrate a clear benefit from the new treatment, the trial may also be stopped early because to continue would be unethical. In addition, if the differences in primary response variables are so unimpressive that the prospect of a clear result is extremely unlikely, it may not be justifiable in terms of time, money and effort to continue the trial. Finally, monitoring of response variables can identify the need to collect additional data in order to clarify questions of benefit or toxicity that may arise during the trial. In order to fulfill the monitoring function, the data must be collected and processed in a timely fashion as the trial progresses.

While the process of repeated significance testing is required for ethical, scientific and

1) This paper was supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1996.
2) Associate Professor, Department of Statistics, Korea University, Seoul, 136-701, Korea.

economic reasons, problems of a statistical nature are raised. (cf. Armitage, McPherson and Rowe, 1969; McPherson and Armitage, 1971). If repeated statistical tests are performed on accumulating data at a conventional significance level (α=0.05), the overall Type I error rate will be substantially higher than intended. For example, the actual type I error would escalate to 8% for 2 such analyses and 14% for 5 analyses.

In order to avoid the problem of having such a substantial risk of a false positive result, various sequential testing procedures have been employed (cf. Armitage, 1975). However, the sequential models assume that the data will be tested after each pair of subjects or after each outcome. In fact, many multicenter studies have a data monitoring committee which meets at regularly scheduled intervals so that decisions to continue or terminate a trial are usually not made after each event or pairs of events. It would be more practical to consider methods appropriate for a number of planned analyses being conducted at specified intervals. This led to the development of group sequential methods and their practical applications to common clinical trials (cf. Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983). These procedures are briefly reviewed in section 2, and more details can be found in DeMets (1987) and Lee (1994-a). In many chronic disease clinical trials, the major endpoint of interest is time to an event (death, relapse etc.). Generally, in such trials, patients enter the study during some staggered entry accrual period and the final analysis is conducted after a predetermined follow-up period. Usually at the final analysis not all the events are observed and thus we have censored survival data. These studies also monitor for evidence of benefit or toxicity, and thus are subject to the repeated testing problem. Many authors proposed the group sequential methods for analyzing survival data (cf. Breslow, 1969; Jones and Whitehead, 1979; Gail, DeMets and Slud, 1981; Tsiatis, 1982; Slud and Wei, 1982; Sellke and Siegmund, 1983; Slud, 1984; DeMets and Gail, 1985; Tsiatis, Rosner and Tritchler, 1985). Most of these methods, however, assume that the failure time distribution depends on the treatment variable and no other external factors. In 1990's, there have been much progress in the development of sequential monitoring methods for comparing survival distributions in clinical trials. I briefly review these recently proposed group sequential methods for survival analysis in section 3. Included are ten methods proposed by Lan and Lachin (1990), Lin (1991, 1992), Gu and Ying (1993), Keaney and Wei (1994), Lan, Rosenberger and Lachin (1995), Lee and Sather (1995), Tsiatis, Boucher and Kim (1995), Lin, Shen, Ying and Breslow (1996) and Betensky(1997). By comparing the contributions and the restrictions of these methods, I provide the guidelines for choosing the appropriate sequential methods in a given situation.

A fixed duration study accrues patients during the accrual period, and the patients are followed during the follow-up period until sufficient number of events of interest has been gathered to test the null hypothesis of no treatment difference. In most clinical trials, fixed-duration designs are being used even when the accumulating data are analyzed periodically. If a sufficiently large differences were found between treatments, early termination of the trial should be considered. The design aspects regarding accrual and follow-up periods have been discussed (cf. Kim and Tsiatis, 1990; Kim, 1992; Kim, Boucher and Tsiatis, 1995)

and these discussions are briefly reviewed in section 4.

2. Group Sequential Testing

Instead of pairing individual patients, the group sequential procedure compares groups of patients accrued during the same period of time. Suppose that we consider K interim analyses in comparing two or more treatment effects. Let $S(k)$ be the standardized summary statistic at the k-th interim analysis $(k=1,\cdots,K)$, e.g. the mean treatment difference divided by its standard error. We plan to stop the trial at k-th interim analysis if $|S(k)|$ exceeds a chosen boundary value b_k. The probabilities π_k $(1\leq k\leq K)$ are chosen such that

$$\pi_1+\cdots+\pi_K=\alpha, \tag{2.1}$$

where α is the type I error and, under the null hypothesis,

$$P_0\{|S(1)|\leq b_1,\cdots,|S(k-1)|\leq b_{k-1},|S(k)|>b_k\}=\pi_k. \tag{2.2}$$

Then it follows that $P_0\{|S(k)|>b_k$ for *some* $1\leq k\leq K\}=\alpha$. If $(S(1),\cdots,S(k))$ follows asymptotically a multivariate normal distribution for any $k(\leq K)$ and its covariance matrix can be consistently estimable, then the sequential boundaries $\{b_k, k=1,\cdots,K\}$, which satisfy the above two conditions, can be constructed recursively using for example the MULNOR program (cf. Schervish, 1984).

Slud and Wei (1982) suggested that the probabilities $\pi_1,\cdots,\pi_K$, summing to α, be prespecified. Fleming, Harrington and O'Brien (1984) have extended the Slud-Wei procedure to allow for modification of the choice of K during the course of the trial. Lan and DeMets (1983) introduced an increasing 'error spending rate' function $\alpha^*(t)$, with $\alpha^*(0)=0$ and $\alpha^*(1)=\alpha$, which allocates the amount of type I error that can be spent at each interim analysis, and set $\pi_k=\alpha^*(t_k)-\alpha^*(t_{k-1})$. Here, t_k is called the 'information time' at the k-th interim analysis, and is defined as the proportion of the total information from the study that has accrued by the k-th analysis. In both methods, the boundary values $b_1, b_2,\cdots, b_K$ can be sequentially constructed as the actual group sizes are observed. Thus, independent increments of the interim test statistics are not required. Note also that the boundary value b_k depends only on the sizes of the current and past groups as well as the values $\pi_1,\cdots,\pi_k$. The Lan-DeMets method allows for changes in frequency and spacing of interim analyses, but the maximum amount of information needs to be estimated at the start of the trial. Proschan,

Follmann and Waclawiw (1992) have studied the effects of assumption violations on type I error inflation in the different group sequential procedures, and have shown that changes in future monitoring times may substantially inflate the overall type I error with the Slud-Wei or Fleming-Harrington-O'Brien approach, but only a little with the Lan-DeMets spending function approach.

If each individual has only a single response, then the test statistic $S(k)$ at the k-th interim analysis behaves like the partial sum of the independent random variables. Hence, the successive test statistics follow asymptotically a multivariate normal distribution with independent increments and the form of their covariance matrix is also simplified. A simple algorithm, developed by Reboussin, DeMets, Kim and Lan (1992), using recursive numerical integration can be used to obtain the sequential boundaries that satisfy (2.1) and (2.2). This algorithm implements a framework provided by Lan and DeMets(1983) and allows for flexible and unequally spaced interim analyses.

Assuming the independent and equal increments, Pocock (1977) suggested setting $b_1 = \cdots = b_K = c_P$, and O'Brien and Fleming (1979) set $b_k = c_B\sqrt{\frac{K}{k}} (1 \le k \le K)$. The constants c_F and c_B depend on K and α. Their computation is performed using the recursive numerical integration by Armitage, McPherson and Rowe (1969). Wang and Tsiatis (1987) considered a more general class of boundaries which includes both Pocock and O'Brien-Fleming boundaries. According to Lan and DeMets (1983), with $t_k = \frac{k}{K}$, $\alpha^*(t) = \alpha \log\{1 + (e-1)t\}$ gives sequential boundaries similar to those of Pocock (1977) and $\alpha^*(t) = 2\left\{1 - \Phi\left(z_{\frac{\alpha}{2}}/\sqrt{t}\right)\right\}$, where Φ is the standard normal distribution function and $z_{\frac{\alpha}{2}}$ is chosen so that $\Pr(z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$, gives sequential boundaries similar to those of O'Brien and Fleming (1979). Design considerations for the choice of the error spending rate function have been discussed by Kim and DeMets (1987) and Jennison and Turnbull (1989).

When each patient has repeated measurements or another kind of multivariate observations, there is no guarantee that the interim statistics have independent increments. Therefore, the asymptotic multivariate normality of successive test statistics and their covariance matrix need to be directly derived to construct the sequential boundaries, satisfying (2.1) and (2.2). Recently some procedures have been proposed for conducting interim analyses with multivariate observations, and nine methods among them were discussed in detail by Lee (1994-b).

3. Sequential Monitoring of Survival Data

3.1. Classical sequential methods

In many clinical trials, we are interested in comparing the survival distribution of different treatment groups. Generally, in such trials, patients enter the study during some staggered entry accrual period and the final analysis is conducted after a predetermined follow-up period. In these studies, early termination of the clinical trial should also be considered if large treatment differences occur.

The group sequential methods described in the previous section require that the interim test statistics $(S(1),\cdots,S(K))$ are asymptotically multivariate normal. The recursive construction of the boundary values at interim analyses is based on this distribution theory. Hence, for a survival study, many authors have tried to derive the asymptotic distribution of the interim test statistics. Many researchers have shown that the numerator of the sequentially computed logrank test behave like a partial sum of independent normal random variables with variance that grows proportionately to the number of failures. When the patients enter the trial sequentially and their response times are subject to random censoring, Tsiatis (1982) derived the asymptotic joint distribution of the score process for Cox's proportional hazards model (cf. Cox, 1972), and Sellke and Siegmund (1983) showed that such a process in fact converges weakly to a time-changed Brownian motion. Slud and Wei (1982) have shown that the modified Wilcoxon (Gehan) statistics follow the asymptotically multivariate normal distribution but have dependent increments. Gail, DeMets and Slud (1981) also showed by simulation that for small samples the group sequential procedure assumptions can be violated using the logrank statistics but still the overall type I error seems to be fairly robust. Tsiatis, Rosner and Tritchler (1985) have used the proportional hazards model introduced by Cox (1972) to allow the dependence of failure time upon other covariates. They have shown that the sequentially computed score test statistic, derived from a partial likelihood (cf. Cox, 1975), also follows asymptotically a multivariate normal distribution.

3.2. Recently proposed methods

Very recently, in 1990's, there have been much progress in the development of sequential monitoring methods for comparing survival distributions in clinical trials. In this section, I briefly review ten recently proposed group sequential methods for survival analysis: Lan and Lachin (1990), Lin (1991, 1992), Gu and Ying (1993), Keaney and Wei (1994), Lan, Rosenberger and Lachin (1995), Lee and Sather (1995), Tsiatis, Boucher and Kim (1995), Lin, Shen, Ying and Breslow (1996) and Betensky(1997).

Lan and Lachin

Lan and Lachin (1990) described a new estimate of the information time, which is defined as the fraction of total information available at the time of interim analysis, and used the

estimated information time to implement the group sequential logrank tests in a maximum duration trial. Total information can be expressed by the number of events to be accrued if the logrank test is employed. In a maximum duration trial, in which the trial ends when a fixed period of calendar time has elapsed, the total number of events accrued is unknown at the time of interim analysis and the information times need to be estimated. Lan and Lachin proposed to use the fraction of total patient exposure as a convenient surrogate measure of information, leading to conservative group sequential boundaries. They also investigated by simulation the properties of the sequential boundaries based on this estimator of information time.

Lin

Lin (1991) has proposed a sequential testing procedure with multiple time-to-event endpoints. The proposed test statistic at each interim analysis is a weighted sum of the linear rank statistics from each endpoint with respect to the marginal distributions of the multiple endpoints. This class of linear rank statistics contains logrank, Peto-Prentice-Wilcoxon, Gehan-Wilcoxon, and more generally the G^{ρ} statistics of Harrington and Fleming (1982). The weights can be chosen to maximize asymptotic power against a specified alternative (cf. Wei and Johnson, 1985). The multivariate asymptotic normality of the test statistic is derived using the same technique as in Wei and Lachin (1984), and the recursive construction of the boundary values at interim analyses, which satisfy (2.1) and (2.2), is based on this distribution theory. This method also works well with the accelerated failure time model (cf. Kalbfleisch and Prentice, 1980) as well as the proportional hazards model. An application is given in Lin and Wei (1991).

Lin

Lin (1992) proposed a group sequential test of no treatment difference which adjusts for other covariates than the treatment variable with accelerated failure time model, which relates covariates linearly to the logarithm of the failure time. The proposed method can be considered as a useful alternative to the sequential method by Tsiatis, Rosner and Tritchler (1985) because Tsiatis-Rosner-Tritchler method adjusts for other covariates with the Cox proportional hazards model. Actually, these two methods are the same except that the test statistic is calculated on a 'transformed' time scale. This accelerated failure time model is especially appealing to medical investigators due to its straightforward interpretation, and there exist semiparametric efficient parameter estimators for this model.

Gu and Ying

Gu and Ying (1993) proposed a Buckly-James-type score process for repeated significance tests of regression hypothesis with staggered entry data under general right censorship. Their research was motivated by that Buckly-James method (1979) is an appropriate extension of the least squares method and is a valuable alternative to Cox's partial likelihood method for regression data especially when the accelerated failure time model is more appropriate than the proportional hazards model. By using the martingale theory of the counting process, they have shown that the score process is asymptotically equivalent to a multidimensional Gaussian process with independent increments. The recursive construction of the boundary values at interim analyses, which satisfy (2.1) and (2.2), is based on this distribution theory. The score process can be interpreted as a weighted comparison of transformed survival times, and this method is especially suitable for the accelerated failure time regression model. Through simulation studies, they have found that the proposed test is superior when the underlying error distribution is normal, and the log-rank method is superior when the error distribution is extreme value.

Keaney and Wei

Keaney and Wei (1994) proposed to use differences or ratios of estimated median survival times repeatedly during the two-arm survival study. They presented a repeated confidence interval estimation procedure for the difference or ratio of median survival times for two treatment groups. It can be considered as a generalization of the method by Jennison and Turnbull (1985), which constructs repeated confidence intervals for the median survival time of a single group of patients by inverting a series of sign tests, to the case for testing the equality of two medians. They have demonstrated that sequentially computed differences of two median survival times follows asymptotically multivariate normal distribution, and used a simple resampling method (cf. Parzen, Wei and Ying, 1994) to estimate the asymptotic covariance matrix without involving any subjective nonparametric functional estimate. Boundary values at interim analyses are recursively constructed to satisfy (2.1) and (2.2). The proposed method is purely nonparametric like the sequential logrank test and can be easily modified to obtain repeated confidence intervals for other quantiles.

Lan, Rosenberger and Lachin

Lan, Rosenberger and Lachin (1995) have described sequential monitoring of data using the Peto-Peto-Prentice Wilcoxon statistic (cf. Peto and Peto, 1972, Prentice, 1978) in a maximum duration trial. This statistic, like logrank statistic, has nice property that sequentially computed statistics form a Brownian motion process, but the information is not easily expressible in terms of number of events. They provided guidelines for estimating the information fraction in

a maximum duration trial when this statistic is employed. When there is a relatively low event rate or the survival time is approximately exponential, they recommended estimating the information time as though the logrank statistic were used.

Lee and Sather

Lee and Sather (1995) presented both parametric and nonparametric sequential testing procedures for clinical trials where the main interest is in testing equality of the cured proportions between two treatments. The parametric procedure is based on the mixture model, described by Farewell (1982), which assumes a logistic model for the cured proportion and a Weibull model for the failure distributions among those who are not cured. The nonparametric procedure is based on the optimal linear rank test, proposed by Gray and Tsiatis (1989), which uses the inverse of the left-continuous version of the pooled Kaplan-Meier estimator as the weighting function among the weighted logrank tests. In both procedures, construction of the sequential boundaries is based on the asymptotic multivariate normality of the sequentially computed test statistics and their covariance matrix. These methods are especially useful in the pediatric cancer clinical trials where there are excellent therapeutic results in a number of different malignancies. They emphasized that either of these tests should be considered only if there is a reasonable a priori evidence of belief that stable plateaus in the survival curves would occur.

Tsiatis, Boucher and Kim

Tsiatis, Boucher and Kim (1995) have recently derived the joint distribution of sequentially computed score tests and maximum likelihood estimates for general parametric models of the survival distribution, where the data are subject to censoring and staggered entry. They represented the sequentially computed score test as a stochastic integral of a counting process martingale (cf. Fleming and Harrington, 1991), and their sequential tests and estimates have an independent increments structure. This result allows for using the general group sequential methodologies discussed in section 2, and these methodologies can be used immediately for any parametric model of the survival distribution. A simulation study is also included to illustrate how these methods work with moderate sample sizes. In the simulation, they applied Lan-DeMets method with both O'Brien-Fleming-type and Pocock-type use function, and found that the empirical results are very close to the expected results for both the score test and the Wald test.

Lin, Shen , Ying and Breslow

Lin, Shen, Ying and Breslow (1996) proposed a sequential method based on the Kaplan-Meier estimator. They have shown that the Kaplan-Meier estimators for the survival

function (or the Nelson-Aalen estimators for the cumulative hazard function) calculated at different calendar time points follow the asymptotically multivariate normal distribution. Although this method is designed for the sequential testing in one-sample case, it is straightforward to extend this result to the two-sample case where the differences between two estimators at each interim analysis are used. Together with the methods by Lee and Sather (1995), the proposed method is useful when a long-term survival rate is the parameter of primary interest. In comparison with the Lee-Sather methods, the proposed method has advantages that we do not parameterize the failure time distribution or impose the proportional hazards structure.

Betensky

Betensky (1997) proposed a simple sequential procedure for comparing survival data from three treatments with the goal of eventually identifying the best treatment. She has applied the methods developed for the sequential analysis for the three treatment groups with normal responses (cf. Siegmund, 1993) to the more complicated case of censored survival data. The procedure consist of two stages of testing. The first test is a global test for detecting on overall treatment effect. If a treatment effect is detected, the worst treatment is eliminated and the second sequential test attempts to identify the better of the two remaining treatments. She has shown that the interim test statistics behave like the sum of two-dimensional standard Brownian motion, and this result allows for the sequential procedures described in section 2 for normal responses to be immediately applied to survival data. Simulation study was conducted to see the performance of the procedure and assess its robustness against departures from the assumption of local alternatives. Design considerations were also discussed.

3.3. Discussion

The group sequential methods described in section 2 require the interim test statistics $(S(1),\cdots,S(K))$ follow the asymptotically a multivariate normal distribution. The above ten methods derived the sequential version of the different types of test statistics under various situations, but all of them used the asymptotically multivariate normality of the interim test statistics and their consistently estimated covariance matrix. In most cases, martingale theory of the counting process made a great contribution to the derivation of the multivariate normality (cf. Fleming and Harrington, 1991). That is, the interim test statistics were represented by the sum of identically independently distributed random variables and the multivariate central limit theorem was applied to derive the multivariate normality. The boundary values at interim analyses, which satisfy (2.1) and (2.2), can be recursively constructed based on this multivariate normality.

It would be worthwhile to summarize each of the above methods according to the type of interim statistics (cf. Table 1). It can provide the guidelines for choosing the appropriate methods in a given situation since each test statistic has its own strong and weak points.

The accelerated failure time model can be considered as a useful alternative to the Cox's proportional hazards model. It is especially appealing to medical investigators due to its straightforward interpretation. When the accelerated failure time model is more appropriate, the readers need to consider using the methods by Lin (1992) and Gu and Ying (1993). The method by Lin (1991) also works well with the accelerated failure time model as well as the proportional hazards model.

Table 1. Interim test statistics used in the methods

Test statistics	Methods	Characteristics
weighted logrank statistics	Lan and Lachin (1990) Lin (1991) Lan, Rosenberger and Lachin (1995) Lee and Sather (1995)	logrank G^{ρ} class Peto-Prentice Gray-Tsiatis cure statistics
score process	Lin (1992) Gu and Ying (1993) Lee and Sather (1995) Tsiatis, Boucher and Kim (1995) Betensky (1997)	accelerated failure time model Buckley-James type Farewell mixture model general parametric model three treatment groups
survival estimates	Keaney and Wei (1994) Lin, Shen, Ying and Breslow (1996)	median survival times cumulative hazards estimates (or Kaplan-Meier estimates)

In some clinical trials situations, one may find that a large portion of the patients never experience the adverse outcome event of interests, and there appears to be no evidence of that event occurring after a certain period of follow up. One area where this occurs frequently is in clinical trials of pediatric cancer where there are excellent therapeutic results in a number of different malignancies. When the main interest is in the comparison of two cure rates, the methods by Lee and Sather (1995) and Lin, Shen, Ying and Breslow (1996) are useful for sequential testing. Table 2 also summarizes the background and the applications of each method.

Table 2. Background and applications of the methods

Methods	Background and/or Applications
Lan-Lachin (1990)	- based on logrank statistic - appropriate for proportional hazards model
Lin (1991)	- based on G^{ρ} class - allows for multiple endpoints - appropriate for proportional hazards model as well as accelerated failure time model
Lin (1992)	- appropriate for accelerated failure time model - appealing due to straightforward interpretation
Gu-Ying (1993)	- based on Buckley-James type statistic - appropriate for accelerated failure time model
Keaney-Wei (1994)	- based on median survival times - estimate repeated confidence interval
Lan-Rosenberger-Lachin (1995)	- based on Peto-Prentice ststistics - estimate information fraction time
Lee-Sather (1995)	- test equality of cure rates - based on Farewell mixture model (parametric) - based on Gray-Tsiatis cure statistics (nonparametric)
Tsiatis-Boucher-Kim (1995)	- based on the score statistic - allows for general parametric model
Lin-Shen-Ying-Breslow (1996)	- test equality of long-term survival rates - based on the cumulative hazards estimates or Kaplan-Meier estimates
Betensky (1997)	- compares three treatment groups - based on two-stage testing

4. Design of Sequential Survival Clinical Trials

The issues of determining study duration for clinical trials with survival data have been addressed by several researchers (cf. Pasternack and Gilbert, 1971; George and Desu, 1974; Rubinstein, Gail and Santner, 1981; Lachin, 1981; Freedman, 1982). Recently, Park, Kim and Lee (1997) have reviewed the existing literature concerning commonly used sample size formulae in the design of randomized clinical trials with survival endpoints, and compared the assumption, the power and the sample size calculation of these methods. They also compared

by simulation the expected power and the observed power of each method under various circumstances, and provided the guidelines in terms of practical usage.

In many clinical trials, there is a set of well-known prognostic factors that are used for stratification in the randomization scheme to avoid confounding the treatment effect with an imbalance in those factors. Many authors have proposed sample size formulae for this case (cf. Bernstein and Lagakos, 1978; Schoenfeld; 1983; Palta and Amini, 1985; Lachin and Foulkes, 1986). Kim, Park and Lee (1997) have also reviewed and compared these methods when the stratification should be considered.

While there have been considerable attention to the design of fixed duration survival studies, only a few authors have paid attention to the design of sequential clinical trials with failure time data. The study duration in clinical trials with failure time data consists of two periods: an accrual period during which patients are entered serially and the follow-up period during which no patients are entered, but those who already entered are followed until failures occur or until the time of study termination, subject to random loss to follow up. When we refer to a group sequential design for a clinical trial with failure time data, we mean the specification of (a) the use function, (b) the durations of accrual and follow-up periods, and (c) the number and times of interim analyses. Although the use function approach as described by Lan and DeMets (1983) maintains the desired significance level without having to specify the number and times of interim analyses, these interim times need to be specified because the number and times of interim analyses have some effect on the power and early stopping properties of these sequential methods.

In a maximum duration trial, in which the trial ends when a fixed period of calendar time has elapsed, the total number of events accrued is unknown at the time of interim analysis and the information times need to be estimated. Under the maximum duration trial, Kim and Tsiatis (1990) proposed a unified design procedure based on the use function approach by Lan and DeMets (1983) and the fixed duration design by Rubinstein, Gail and Santner (1981). The have also compared different group sequential designs in terms of various expected stopping times, and provided guidelines for selecting appropriate design specifications. Kim (1992) extended the Kim-Tsiatis procedure to allow for the stratified randomization and the unequal patient allocation between treatments. His method is based on the use function approach and the fixed duration design by Bernstein and Lagakos (1978). While the above two methods are proposed for a maximum duration trial, Kim, Boucher and Tsiatis (1995) proposed a design procedure for maximum information trials, in which the maximum amount of information is fixed, and investigated the properties of maximum information trials for different group sequential boundaries. They also illustrated how to maintain the type I error in maximum duration trials by using correlation structure of the sequentially computed logrank statistics, and compared maximum information trials and maximum duration trials.

5. Concluding Remarks

Before 1990, most of the group sequential methods for comparing survival distributions assume that the failure time distribution depends on the treatment variable and no other external factors. Nonparametric methods are also based on the typical logrank statistics.

In 1990's, however, many methods have been proposed for comparing survival distributions with various types of test statistics such as median survival times, cumulative hazards estimates, Buckley-James type statistics and the G^{ρ} class of logrank statistics. Some methods also allow for many types of models like accelerated failure time model, mixture model and the general parametric models as well as the proportional hazards model. As discussed in section 3, each of the test statistics and the models has its own strong and weak points, and thus the biostatisticians should be careful to choose the most appropriate methods in their clinical trial.

While there have been much progress in the group sequential testing procedures, only a few researchers have paid attention to the group sequential design of the survival clinical trial. Most of the above testing procedures are also based on the maximum duration trial, in which the trial ends when a fixed period of calendar time has elapsed. Development of the group sequential designs of the maximum information trial and the testing procedures based on the maximum information trial could be a possible future research.

References

[1] Armitage, P. (1975). *Sequential Medical Trials*, 2nd edition, John Wiley and Sons, New York.

[2] Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance tests on accumulating data, *Journal of the Royal Statistical Society*, Ser. A, 132, 235-244.

[3] Bernstein, D. and Lagakos, S. W. (1978). Sample size and power determination for stratified clinical trials, *Journal of Statistical Computing and Simulation*, 8, 65-73.

[4] Betensky, R. A. (1997). Sequential analysis of censored survival data from three treatment groups, *Biometrics*, 53, 807-822.

[5] Breslow, N. (1969). On large sample sequential analysis with applications to survivalship data, *Journal of Applied Probability*, 6, 261-274.

[6] Buckley, J. and James, I. (1979). Linear regression with censored data, *Biometrika*, 66, 429-436.

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

[8] Cox, D. R. (1975). Partial likelihood, *Biometrika*, 62, 269-276.

[9] DeMets, D. L. (1987). Practical aspects in data monitoring: A brief review, *Statistics in Medicine*, 6, 753-760.

[10] DeMets, D. L. and Gail, M. H. (1985). Use of logrank tests and group sequential methods at fixed calendar times, *Biometrics*, 41, 1039-1044.

[11] Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics*, **38**, 1041-1046.

[12] Fleming, T. R., and Harrington, D. P. (1991). *Counting Process and Survival Analysis*, Wiley, New York.

[13] Fleming, T. R., Harrington, D. P., and O'Brien, P. C. (1984). Designs for group sequential tests, *Controlled Clinical Trials*, 5, 348-361.

[14] Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using the logrank test, *Statistics in Medicine*, 1, 121-129.

[15] Gail, M. H., DeMets, D. L., and Slud, E. V. (1981). Simulation studies on increments of the two-sample logrank score test for survival data, with application to group sequential boundaries. In: *Survival Analysis*, IMS Lecture Notes Monograph Series 2 (Johnson and Crowley Eds.). Hayward, California.

[16] George, S. L. and Desu, M. M. (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases*, 27, 15-24.

[17] Gray, R. J. and Tsiatis, A. A. (1989). A linear rank test for use when the main interest is in differences in cure rates, *Biometrics*, **45**, 899-904.

[18] Gu, M. and Ying, Z. (1993). Sequential analysis for censored regression data, *Journal of the American Statistical Association*, 88, 890-898.

[19] Harrington, D. P., and Fleming, T. R. (1982). A class of rank test procedures for censored survival data, *Biometrika*, 69, 553-566.

[20] Jennison, C., and Turnbull, B. W. (1985). Repeated confidence intervals for median survival time, *Biometrika*, 72, 619-625.

[21] Jennison, C., and Turnbull, B. W. (1989). Interim analyses: The repeated confidence interval approach, *Journal of the Royal Statistical Society*, Ser. B, 51, 305-361.

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

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

[24] Keaney, K. M. and Wei, L. J. (1994). Interim analyses based on median survival times, *Biometrika*, 81, 279-286.

[25] Kim, K. (1992). Study duration for group sequential clinical trials with censored survival data adjusting for stratification, *Statistics in Medicine*, 11, 1477-1488.

[26] Kim, K. and Tsiatis, A. A. (1990). Study duration and power consideration for clinical trials with survival response and early stopping rule, *Biometrics*, 46, 81-92.

[27] Kim, K., Boucher, H. and Tsiatis, A. A. (1995). Design and analysis of group sequential logrank tests in maximum duration versus information trials, *Biometrics*, 51,

988-1000.

[28] Kim, S. W., Park, M. and Lee, J. W. (1997). Sample size determination for stratified survival studies, Unpublished manuscript.

[29] Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials, *Controlled Clinical Trials*, 2, 93-113.

[30] Lachin, J. M. and Foulkes, M. A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification, *Biometrics*, 42, 507-519.

[31] Lan, K. K. G., and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials, *Biometrika*, 70, 659-663.

[32] Lan, K. K. G. and Lachin, J. M. (1990). Implementation of group sequential logrank tests in a maximum duration trial, *Biometrics*, 46, 759-770.

[33] Lan, K. K. G., Rosenberger, W. F. and Lachin, J. M. (1995). Sequential monitoring of survival data with the Wilcoxon statistic, *Biometrics*, 51, 1175-1183.

[34] Lee, J. W. (1994-a). Group sequential testing in clinical trials with multivariate observations: A review, *Statistics in Medicine*, 13, 101-111.

[35] Lee, J. W. (1994-b). An overview of group sequential procedures, *The Korean Journal of Applied Statistics*, 7, No.2, 35-51.

[36] Lee, J. W., and Sather, H. N. (1995). Sequential methods for comparison of cure rates in clinical trials, *Biometrics*, 51, 756-763.

[37] Lin, D. Y. (1991). Nonparametric sequential testing in clinical trials with incomplete multivariate observations, *Biometrika*, 78, 123-131.

[38] Lin, D. Y. (1992). Sequential logrank tests adjusting for covariates with covariates with the accelerated life model, *Biometrika*, 79, 523-529.

[39] Lin, D. Y., Shen, L., Ying, Z. and Breslow, N. E. (1996). Group sequential designs for monitoring survival probabilities, *Biometrics*, 52, 1033-1041.

[40] Lin, D. Y. and Wei, L. J. (1991). Repeated confidence intervals for a scale change in a sequential survival study, *Biometrics*, 47, 289-294.

[41] McPherson, C. K., and Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true, *Journal of the Royal Statistical Society*, Ser.A, 134, 15-25.

[42] O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials, *Biometrics*, 35, 549-556.

[43] Palta, M. and Amini, S. (1985). Consideration of covariates and stratification in sample size determination for survival time studies, *Journal of Chronic Diseases*, 38, 801-809.

[44] Park, M., Kim, S. W., and Lee, J. W. (1997). Sample size determination for comparative survival studies, *Journal of Applied Statistics*, To appear.

[45] Parze, M. I., Wei, L. J. and Ying, Z. (1994). A resampling method based on pivotal estimating functions, *Biometrika*, 81, 341-350.

[46] Pasterack, B. S. and Gilbert, H. S. (1971). Planning the duration of long-term survival time studies designed for accrual by cohorts, *Journal of Chronic Diseases*, 24, 681-700.

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

[48] Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrika*, 64, 191-199.

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

[50] Proschan, M. A., Follmann, D. A., and Waclawiw, M. A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring, *Biometrics*, 48, 1131-1143.

[51] Reboussin, D. M., DeMets, D. L., Kim, K. and Lan, K. K. L. (1992). Programs for computing Lan-DeMets bounds, Technical Report 60, University of Wisconsin-Madison, Biostatistics Center.

[52] Rubinstein, L. V., Gail, M. H. and Santner, T. J. (1981). Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation, *Journal of Chronic Disease*.

[53] Schervish, M. J. (1984). Multivariate normal probabilities with error bound (with corrections in 1985), *Applied Statistics*, 33, 81-94.

[54] Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model, *Biometrics*, 39, 499-503.

[55] Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazard model, *Biometrika*, 70, 315-326.

[56] Siegmund, D. (1993). A sequential clinical trial for comparing three treatments, *Annals of Statistics*, 21, 464-483.

[57] Slud, E. V. (1984). Sequential linear rank tests for two-sample censored survival data, *Annals of Statistics*, 12, 551-571.

[58] Slud, E. V., 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.

[59] 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.

[60] Tsiatis, A. A., Boucher, H, and Kim, K. (1995). Sequential methods for .parametric survival models, *Biometrika*, 82, 165-173.

[61] Tsiatis, A. A., Rosner, G. L. and Tritchler, D. L. (1985). Group sequential tests with censored survival data adjusting for covariates, *Biometrika*, 72, 365-373.

[62] Wang, S. K., and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential approach, *Biometrics*, 43, 193-199.

[63] Wei, L. J., and Johnson, W. E. (1985). Combining dependent tests with incomplete repeated measurements, *Biometrika*, 72, 359-364.

[64] Wei, L. J., and Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations, *Journal of the American Statistical Association*. 79. 653-661.