응용통계연구 제 7권 2호
1994년 9월 pp. 35-51

An Overview of Group Sequential Procedures

Jae Won Lee[1)]

Abstract

For ethical and economic reasons, clinical trials must be repeatedly monitored for evidence of treatment benefit or harm. Repeated testing at conventional critical values can substantially inflate the overall type I error rate. To maintain acceptable levels, group sequential procedures have been developed for clinical trials. This article gives a brief overview of the group sequential procedures.

1. Introduction

During the past four decades, the randomized control clinical trial has emerged as an important tool for the careful evaluation of new drugs or therapeutic procedures. Ethical considerations require that investigators should not deploy patient resources inefficiently or unnecessarily. Thus a trial that shows early benefit or unexpected toxicity may be seriously considered for early termination. To achieve this, statisticians conduct interim statistical analyses periodically on accumulating data.

While the interim analyses are required for ethical, scientific and economic reasons, the process of repeated testing seriously inflates the overall type I error (cf. Armitage, McPherson and Rowe, 1969; McPherson and Armitage, 1971). If the conventional critical value of ± 1.96 is used at each interim analysis, for example, the actual type I error would escalate (e.g. 8% for 2 analyses, 14% for 5 analyses, and 25% for 20 analyses).

Various methods have been proposed to achieve the desired levels of type I error (α) and type II error (β). Wald (1947) introduced sequential methods for testing a simple hypothesis versus a simple alternative which needed continual evaluation. After each new sample is obtained, the likelihood ratio test statistic L is computed and the sampling continues as long as $\beta/(1-\alpha) \leq L \leq (1-\beta)/\alpha$. If the ratio is less than the lower bound, the null hypothesis is accepted and if the upper bound is exceeded, the null hypothesis is rejected. This test, called the sequential probability ratio test (SPRT), forms the basis for classical sequential designs where the desired levels of type I and type II error are

1) Department of Statistics, Korea University, Anamdong 5-1, Sungbukku, Seoul, 136-701, KOREA.

achieved. This method was later adapted for medical research as described by Armitage (1975). He has proposed repeated significance testing boundaries derived from testing after every pair of data points, using methodology from earlier work of Armitage, McPherson and Rowe (1969).

Several problems arise in the application of these sequential methods. A concern is that 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. Another limitation of these sequential methods is the requirement of specifying an alternative in order to establish the stopping boundaries.

A simple but significant modification by Pocock (1977) of Armitage's repeated significance testing model led to a group sequential methodology for monitoring boundaries. The essence of the modification was to recognize that data monitoring committees meet at scheduled times at which information on recently enrolled patients is available. Instead of pairing individual patients, the group sequential procedure compares groups of patients accrued during the same period of time. Various group sequential procedures have been proposed (cf. Pocock, 1977; O'Brien and Fleming, 1979; Slud and Wei, 1982; Lan and DeMets; 1983). These procedures are briefly reviewed in section 2, and more details can be found in DeMets (1987) and Jennison and Turnbull (1989).

Most work related to group sequential testing assumes the sequentially computed test statistics have independent increments. This assumption is no longer reasonable when subjects enter the trial sequentially and a response variable is measured for each subject at successive follow-up visits. Several parametric and nonparametric procedures have been recently proposed to deal with the repeated measurements in group sequential trials (cf. Armitage, Stratton and Worthington, 1985; Geary, 1988; Wei, Su and Lachin, 1990; Lee and DeMets, 1991, 1992; Wu and Lan, 1992; Su and Lachin, 1992). These methods are briefly reviewed in section 3, and more details were discussed by Lee(1994).

Another point of view that has developed in parallel with the sequential methods is that of curtailed procedures. During the monitoring process of accumulating data, one question which might be asked is whether the current data are so impressive, either for or against the null hypothesis, that the acceptance or rejection of the null hypothesis by the final test statistic is already determined. While the group sequential methods focus on existing data, the curtailed approach focuses on existing data plus data yet to be observed. These approaches are briefly discussed in section 4.

Despite the obvious advantages of the group sequential approaches, it has been claimed that these methods remain underemployed because of inconsistencies (cf. Berry, 1987). In particular, the same data can lead to different conclusions whether there are interim analyses or not. Falissard and Lellouch (1992,1993) proposed some new approaches that

are especially conceived to eliminate the inconsistencies of the previous methods. The new procedures impose a succession of r tests significant at a nominal level very close to α, in order to achieve the overall type I error α. These approaches are briefly reviewed in section 5.

This paper gives a brief overview of some recently developed group sequential procedures. The purpose of this paper is to give readers insight into these methods and to provide guidelines for choosing the appropriate method in a specific situation.

2. Group sequential testing with a single response

Consider the simple case of independent responses that are normally distributed with known variance. Suppose we want to compare two treatment means. We assume the null hypothesis $H_0 : \theta = \mu_1 - \mu_2 = 0$, where μ_1 and μ_2 represent the unknown population means for each of two treatment arms, and the interim analysis occurs after randomization of each independent group of $2n$ patients, n per treatment arm. We define the standardized test statistic at the k-th ($1 \le k \le K$) interim analysis, $S(k)$, as the mean treatment difference (estimated from the data up to the k-th analysis) divided by its standard error. For the j-th group of $2n$ patients, let $\overline{X_{1j}}$ and $\overline{X_{2j}}$ represent the average responses. Then $S(k)$ is given by $S(k) = \sum_{j=1}^{k} Z_j/\sqrt{k}$, where $Z_j = \sqrt{n}(\overline{X_{1j}} - \overline{X_{2j}})/\sqrt{2\sigma^2}$. It will also be convenient to define a quantity $I(k)$ called an "information time". In this case it is defined as $I(k) = 2nk/\sigma^2$, the Fisher information for θ. We also define an unstandardized statistic $S^*(k)$ by $S^*(k) = S(k)\sqrt{I(k)} = S(k)\sqrt{2nk/\sigma^2}$. Under H_0, the sequence $\{S^*(k);\ k = 1, \cdots, K\}$ has a multivariate normal distribution with zero means, variances $2nk/\sigma^2$ and independent increments, i.e. the covariances are $\mathrm{cov}(S^*(k), S^*(j)) = \mathrm{var}(S^*(k)) = 2nk/\sigma^2$, $k < j$. Thus $\{S^*(k),\ k = 1, \cdots, K\}$ can be treated as the values of a standard Brownian motion observed at times $\{I(k),\ k = 1, \cdots, K\}$ in the Brownian motion time scale. Suppose probabilities $\{\pi_k, 1 \le k \le K\}$, are chosen such that

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

then from this knowledge of the joint distribution of $\{S^*(k),\ k=1,\cdots,K\}$, it is possible to construct the boundary values $\{b_k,\ k=1,\cdots,K\}$ recursively, such that under H_0:

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

Then it follows that $P_0\{|S(k)| > b_k;\ \text{for some } 1 \le k \le K\} = \alpha$. The nominal two-sided significance level at the k-th analysis is $2\alpha_k$ where $\alpha_k = 1-\Phi(b_k)$ and Φ denotes the standard normal distribution function.

There have been several suggestions as to how to choose discrete sequential boundaries $\{b_k,\ k=1,\cdots,K\}$. The two best known suggestions are due to Pocock (1977) and O'Brien and Fleming (1979). Pocock (1977) suggested setting $\alpha_1 = \cdots = \alpha_K$ or equivalently $b_1 = \cdots = b_K = Z_P(K,\alpha)$, where $Z_P(K,\alpha)$ is a constant depending on K and α. O'Brien and Fleming (1979) let $b_k = Z_B(K,\alpha)\sqrt{K/k}$ $(1 \le k \le K)$, where again $Z_B(K,\alpha)$ is a constant. This is equivalent to choosing boundary points that are constant on the $S^*(k)$ scale. Here, the constants $Z_P(K,\alpha)$ and $Z_B(K,\alpha)$ are chosen so that (2.1) and (2.2) are satisfied. For the case of error $\alpha = 0.05$, Haybittle (1971) had chosen $b_1 = \cdots = b_{K-1} = 3$, $b_K = 1.96$, the standard 5% point for a fixed sample size test. In this case the left hand side of (2.2) will obviously exceed α, but only by a slight amount. For $K=5$ and $\alpha = 0.05$, a graphical comparison of the Pocock, the O'Brien-Fleming and the Haybittle boundaries is given in Appendix. It is desirable to use the Pocock method when all K interim analyses are equally important and the O'Brien-Fleming method when the interim analyses become increasingly more important.

Slud and Wei (1982) suggested that exit probabilities $\pi_1, \cdots, \pi_K$, summing to α, be prespecified and critical values b_k be sequentially computed by solving (2.1) and (2.2) as the actual group sizes are observed. Their method uses only the current and past group sizes and not the unknown sizes of future groups. They considered the sequential testing for the equality of two survival distributions, and showed that the sequentially computed modified Wilcoxon scores, say, $\{S(k), k=1,\cdots,K\}$ in (2.2), has asymptotically a multivariate normal distribution. Lan and DeMets (1983) introduced an 'error spending rate' function $\alpha^*(t)$, $0 \le t \le 1$, which is nondecreasing with $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$. This function

allocates the amount of type I error that can be used or spent in the Brownian motion time scale $I(k)$. A maximum amount of information, I_{max}, must be specified; for normal observations with variance σ^2, $I_{max} = N_{max}/\sigma^2$, where N_{max} is the maximum possible number of observations. Defining $v_k = I(k)/I_{max} = n(k)/N_{max}$, where $n(k)$ is total number of observations up to the k-th analysis, they set $\pi_k = \alpha^*(v_k) - \alpha^*(v_{k-1})$ and solve sequentially for $b_1, b_2, \cdots, b_k$ in (2.2). Note that here b_k depends only on the current and past group sizes, and there is no need to specify K, the total number of looks, in advance. Although this approach needs an accurate estimation of N_{max} or I_{max} at the start of trial, it is very finely tuned to the actual group sizes. The design consideration for choice of the error spending rate function was also discussed by Kim and DeMets (1987-a). A suggestion by Fleming, Harrington and O'Brien (1984) was the extension of Slud and Wei (1982) procedure to allow for the modification of the choice of the maximum number of analyses during the course of experiment. For example, if K and $\pi_1, \cdots, \pi_K$ have been prespecified at the start of the study but, after the k-th analysis, it is decided to change the maximum number of analyses to K', a new choice of specified exit probabilities $\pi'_{k+1}, \cdots, \pi'_{K'}$ is used where $\sum_{i=k+1}^{K} \pi_i = \sum_{i=k+1}^{K'} \pi'_i$. However, the possibility of abuse and threat to credibility make such modification dangerous in practice (cf. Jennison and Turnbull, 1989). 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.

Pocock (1982) derived the 'optimal' boundaries for selected specifications in the sense of minimizing the average sample number. He looked at all possible boundaries $\{b_k;\ k = 1, \cdots, K\}$, and then found the combinations that minimize the average sample number (ASN) or expected sample size for detecting specified treatment difference. Wang and Tsiatis (1987) considered a class of boundaries indexed by a single parameter Δ and showed that the approximately optimal boundaries within this class are approximately optimal overall. A parametric family of group sequential tests is also proposed by Jennison (1987). These parametric tests can be implemented when group sizes are unequal and unpredictable, and are nearly optimal in the sense that these minimize the expected sample size.

The use of one-sided sequential tests was also considered by DeMets and Ware (1980, 1982) and Jennison (1987). Jennison and Turnbull (1983), Chang and O'Brien (1986) and Kim and DeMets (1987-b) presented methods for deriving a confidence interval for parameter following a group sequential test. Jennison and Turnbull (1984) also recommended the use of repeated confidence intervals.

3. Group sequential testing with repeated measurements

If each individual has only a single response, then the test statistics $S(k)$ 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. When each patient has repeated measurements, however, 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 which satisfy (2.1) and (2.2). Recently, both parametric and nonparametric procedures have been proposed for conducting interim analyses with repeated measurements. The reader is referred to Lee (1994) for more extensive discussion.

3.1. Parametric methods

Armitage et al. (1985) have considered the first-order autoregressive (AR(1)) model for repeated measurements as a basis for interim analysis. In this model, $y_{lik}-\mu_{li}=\rho(y_{li,k-1}-\mu_{li})+\varepsilon_{lik}$, where y_{lik} $(l=1,2;\ i=1,\cdots,n;\ k=1,\cdots,K)$ is the k-th response for subject i in treatment group l, and ε_{ijk} is identically independently distributed $(i.i.d.)$ $N(0,\sigma^2(1-\rho^2))$, μ_{li} is $i.i.d.$ $N(\mu_l,\sigma_b^2)$ and ρ is the serial correlation. They have suggested that intermediate significance levels should be adjusted for the ratio of between-subject to within-subject variance. When AR(1) model is allowed, the effect of negative correlation coefficient is to move the results in the direction of group-sequential testing for independent increments. The effect of positive correlation is the opposite; namely, to reinforce the effect of between-subject variation. Their method needs various strong assumptions. One is that all subjects enter the trial at the same time. It was also assumed that there are no missing data, and the repeated measurements on the subjects are assumed to be recorded at equally spaced times. Finally, the data should be normally distributed.

Geary (1988) used $(\sigma_b^2,\phi,\sigma^2,\kappa)$, or 4-parameter, model which is more general than AR(1)

model. Here, ϕ is a autocorrelation coefficient, and κ determines whether within-subject variances are increasing or decreasing. That is, for $\kappa=1$, the errors are stationary and this model is reduced to AR(1) model. His conclusion is similar to, but more general than, that of Armitage et al. (1985), and the model $(\sigma_b^2,\phi,\sigma^2,\kappa)=(0,0,1,1)$ or $(0,0,1,0)$ corresponds to the usual independent measurements case. However, all the assumptions required in Armitage et al. (1985) are still needed in this method.

Wei, Su and Lachin (1990) have assumed a generalized linear model where the repeated measurements from i-th subject, say y_{ij}, follows an exponential family distribution of the form $f(y_{ij})=\exp\{[y_{ij}\theta_{ij}+a(\theta_{ij})+b(y_{ij})]/\phi\}$. $\mu_{ij}=E(y_{ij})=a'(\theta_{ij})$ is assumed to satisfy $h(\mu_{ij})=x_{ij}'\beta$, where h is known link function. They have used the regression methods of Liang and Zeger (1986) to make inferences about one specific component of β, say β_p, which is the treatment difference, at each interim analysis. It follows that the joint distribution of $\hat{\beta_p}=(\hat{\beta_{p1}},\cdots,\hat{\beta_{pK}})$ can be approximated by a multivariate normal, where $\hat{\beta_{pk}}$, $k=1,\cdots,K$, is estimated from the accumulated data up to the k-th interim analysis. From the knowledge of this joint distribution, the boundary values which satisfy (2.1) and (2.2) were constructed recursively. Their procedures can also be applied to discrete repeated measurements as well as continuous measurements, but are not good for testing the equality of the rates of change. Staggered entry and missing values in the data are also allowed.

Lee and DeMets (1991) have assumed that the repeated measurements follow the linear mixed effects model, proposed by Laird and Ware (1982), which can be easily used for highly unbalanced data (cf. Ware, 1985). In the linear mixed effects model, $y_i=X_i\alpha+Z_i\beta_i+e_i$, where y_i is a $n_i\times 1$ vector of observations for the i-th independent subject, X_i and Z_i are known covariate matrices of order $n_i\times p$ and $n_i\times q$ respectively, α is a $p\times 1$ vector of mean parameters to be estimated. The β_i are independent $q\times 1$ random vectors which are assumed to follow a multivariate Gaussian distribution with mean zero and variance matrix D, and the $n_i\times 1$ vectors of residuals e_i are assumed to be independent normal with mean zero and variance matrix R_i. This model has the 4-parameter model of Geary (1988) as a special case. They have estimated the population parameters using the efficient implementation of the Newton-Raphson algorithm, developed by Lindstrom and Bates (1988), and have shown that the difference in these estimators between two treatment groups follow the asymptotic normal distribution.

As in Wei, Su and Lachin (1990), the recursive construction of the boundary values at interim analyses, which satisfy (2.1) and (2.2), are based on this distribution theory. This method can be applied to compare two rates of change as well as two means, although it can not be applied to discrete repeated measurements. This method does not require any of the assumptions in Geary (1988). In other words, staggered entry and missing values in the data are also allowed. In their example, Lee and DeMets (1991) assumed that the errors are conditionally independent (given the random effects). However, simulation studies by Lee and DeMets (in press) have shown that the independence model usually works well even when a more general error structure is adopted.

Wu and Lan (1992) have used two stage random effects model, proposed by Wu and Bailey (1989), to account for informative censoring. They have also proposed to compare treatment effects between two groups by comparing the areas under the two expected response change curves. When the response curves are linear as a function of time in both groups, this comparison is equivalent to comparing the rates of change in the response variable, as discussed in Lee and DeMets (1991). The asymptotic normality of the interim test statistics $(S(1),S(2),\cdots,S(K))$, derived by Lee and DeMets (1991), can be used to recursively construct the boundary values which satisfy (2.1) and (2.2). This method can enjoy all the applications of the method by Lee and DeMets (1991) since both methods actually use the same model.

3.2. Nonparametric methods

Su and Lachin (1992) have considered the multivariate generalized U-statistic, proposed by Wei and Johnson (1985). They proposed using the test statistic at each interim analysis as the weighted sum of the linear rank statistics with respect to the marginal distributions of the multiple endpoints. The weights can also be chosen to maximize asymptotic power against the alternatives (cf. Wei and Johnson, 1985). They have derived the multivariate asymptotic normality of the this statistic, 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 can be especially useful when the main interest is in testing whether the two population mean profiles are similar.

Lee and DeMets (1992) have proposed a sequential testing procedure based on the linear rank statistics which can be applied to either continuous or discrete repeated measurements. Using the asymptotic normality theory in multivariate linear models, developed by Puri and Sen (1985) and extended by Lee, Reboussin and DeMets (1990) to allow for the missing data which are missing at random, they have derived the joint multivariate normality of sequentially computed linear rank test statistics. As in the other methods described above, the recursive construction of the boundary values at interim analyses, which satisfy (2.1)

and (2.2), is based on this distribution theory. It can also be used to compare two rates of change as well as two means. The nonparametric test based on the multivariate linear rank statistics has a variety of applications. In addition to two-sample problem, it can be applied to one-sample, multi-sample problem and the paired comparison problem (cf. Puri and Sen, 1971). This method can enjoy the same applications.

4. Curtailed procedures

The procedures described thus far evaluate the evidence accumulated without explicit consideration of the data yet to be observed. One procedure that considers both observed and unobserved data is referred as 'curtailed procedure'. Deterministic curtailed procedures have been proposed to allow early termination only when reversal of the existing trend is impossible (cf. Halperin and Ware, 1974; DeMets and Halperin, 1981). Since the concern is with what the test statistic results would be at the end of the trial, repeated testing of the data does not increase the type I error rate beyond α level being used in the final test. However, this extreme conservatism allows very little opportunity for early termination. Lan, Simon and Halperin (1982) relaxed the requirement of certainty and proposed a less conservative stochastic curtailed procedure. They considered early termination when the probability of a trend being reversed is small and discussed the effect of this on Type I and Type II error rates.

More specifically, assume that the null hypothesis H_0 will be tested at the end of a trial at time T, using a statistic $S(T)$. Suppose we observe a trend in favor of rejecting H_0 at time $t<T$. We compute the conditional probability, γ_0, of rejecting H_0 at the scheduled end T, assuming H_0 to be true and given the existing data $S(t)$. If γ_0 is 1, then we have a deterministic procedure. If this is suitably large, we might be willing to assume that the current trend is not going to reverse itself. We obtain similar results for the type II error. For a negative or lack of trend, we compute the conditional probability, γ_1, of not rejecting H_0 at the scheduled end T, even if the specified alternative were true. Here, one can often consider a series of alternative hypotheses before one rejects a possible reversal as unlikely.

Because we are not certain that the results will not change, both type I and type II error rates are slightly inflated by this procedure. Lan, Simon and Halperin (1982) showed that type I error can be bounded by α/γ_0 and the type II error by β/γ_1. If $\gamma_0=0.9$ and $\alpha=0.05$, for example, then $\alpha/\gamma_0=0.056$. Others have recently suggested using a prior on

the alternatives and obtaining the predictive power of a positive result (cf. Spiegelhalter, 1986; Spiegelhalter, Freedman and Blackburn; 1986).

5. Succession procedures

One major criticism of most group sequential methods is that the same data lead to different conclusions whether there are interim analyses or not. The idea of requiring that each of a succession of tests be significant at a particular level of significance for early termination of a clinical trial has been presented.

Canner (1977,1984) considered the problem of monitoring survival in long-term clinical trials. He proposed to conclude a 'statistically significant' treatment effect when the values of the observed test statistics all exceed $Z_{(\alpha^*)}$ for r successive analyses or when the final test statistic is greater than $Z_{(\alpha^*)}$. Falissard and Lellouch (1992) modified Canner's approach. If there is a succession of r tests significant at a nominal level α' very close to α, the trial ends with both rejecting H_0 and a significant test at the α level test at the last analysis. If such a succession never occurs, the trial can stop when the current analysis is not significant at the α' level and it is too close to the end of the trial to obtain the suitable succession of r significant tests. In this situation, the trial ends both with not rejecting H_0 and a nonsignificant test at the α level at the last analysis. For example, suppose that we consider a total of 7 interim analyses and 3 successive tests with overall type I error 0.05. In this case, the nominal level α' for each test is 0.0573 which is very close to α. H_0 is rejected as soon as 3 successive tests are significant, but H_0 is accepted if the 5-th interim test is not significant because we can not obtain 3 successive significant tests. Compared to Canner's method, the succession procedure by Falissard and Lellouch (1992) appears to provide the possibility of terminating the trial earlier at the cost of lower power. If $r=1$, then both succession procedures become Pocock's procedure described in section 2.

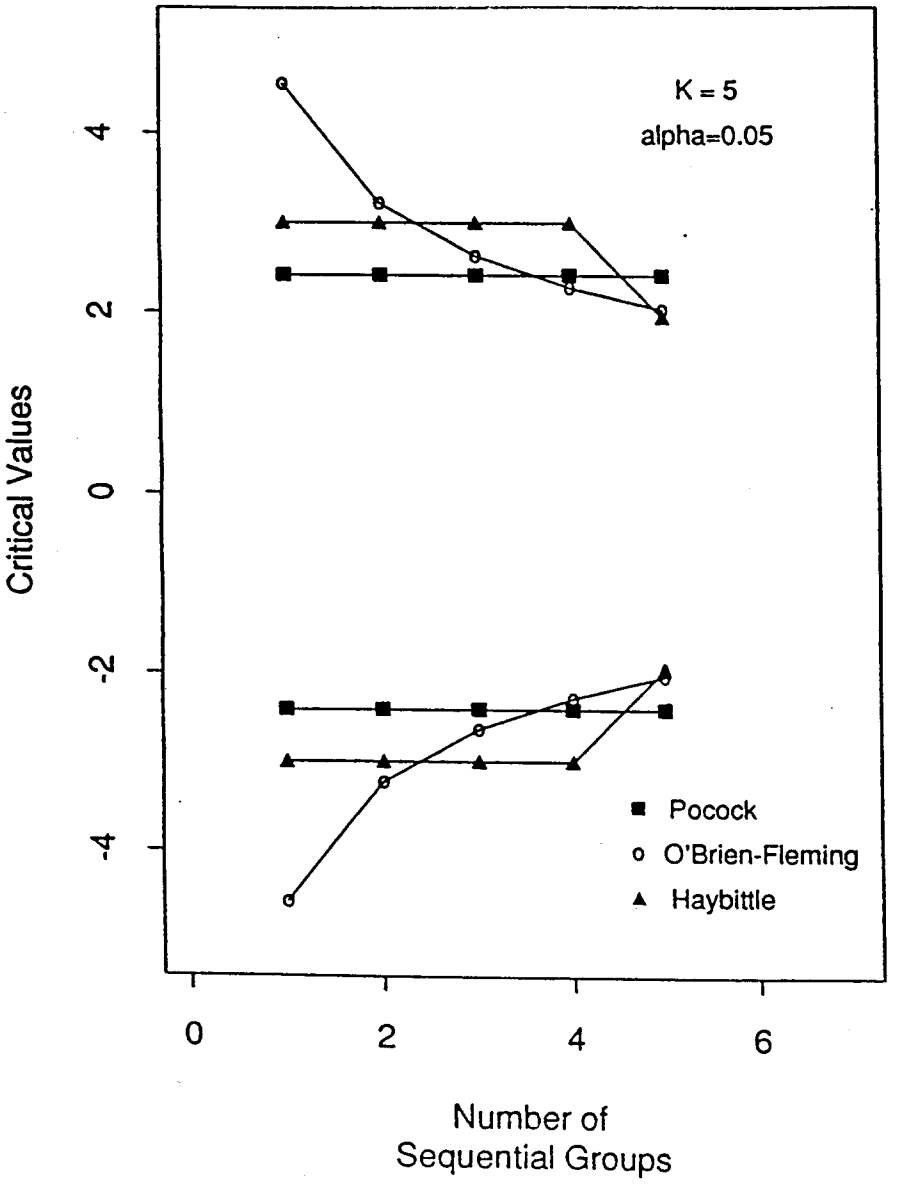
Falissard and Lellouch (1993) extended the above succession procedure in two directions. The first uses the principles of stochastically curtailed tests described in section 4. A lower boundary was added to the original succession procedure to allow the early acceptance of the null hypothesis. They indicated that the new procedure gives the similar power to the original one, but allowing early acceptance of the null hypothesis substantially reduces the average number of patients. The second, adapted to the Lan-DeMets type I

error spending rate function described in section 2, leads to the semisequential analysis which allows flexible times of interim analyses. They also showed that the performances of the semisequential analysis are very close to those of the O'Brien-Fleming method.

Acknowledgements

This research was completed while the author was at the University of Southern California and the Childrens Cancer Group. This work was supported in part by NIH grant CA-13539.

Appendix



A graphical comparison of the boundaries

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, Vol. 132, 235-244.

[3] Armitage, P., Stratton, I. M., and Worthington, H. V. (1985). Repeated significance tests for clinical trials with a fixed number of patients and variable follow-up, *Biometrics,* Vol. 41, 353-359.

[4] Berry, D. A. (1987). Statistical inference, designing clinical trials and pharmaceutical company decisions, *The Statistician,* Vol. 36, 181-189.

[5] Canner, P. L. (1977). Monitoring treatment differences in long-term clinical trials, *Biometrics,* Vol. 33, 603-615.

[6] Canner, P. L. (1984). Monitoring long-term clinical trials for beneficial and adverse treatment effects, *Communications in Statistics-Theory and Methods,* Vol. 13, 2369-2394.

[7] Chang, M. N., and O'Brien, P. C. (1986). Confidence intervals following group sequential tests, *Controlled Clinical Trials,* Vol. 7, 18-26.

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

[9] DeMets, D. L., and Halperin, M. (1981). Early stopping in the two-sample problem for bounded random variables, *Controlled Clinical Trials,* Vol. 3, 1-11.

[10] DeMets, D. L., and Ware, J. H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis, *Biometrika,* Vol. 67, 651-660.

[11] DeMets, D. L., and Ware, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials, *Biometrika,* Vol. 69, 661-663.

[12] Falissard, B., and Lellouch, J. (1992). A new procedure for group sequential analysis, *Biometrics,* Vol. 48, 373-388.

[13] Falissard, B., and Lellouch, J. (1993). The succession procedure for interim analysis: Extensions for early acceptance of H_0 and for flexible times of analysis, *Statistics in Medicine,* Vol. 12, 51-67.

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

[15] Geary, D. N. (1988). Sequential testing in clinical trials with repeated measurements, *Biometrika,* Vol. 75, 311-318.

[16] Halperin, M., and Ware, J. (1974). Early decision in a censored Wilcoxon two-sample

test for accumulating survival data, *Journal of the American Statistical Association,* Vol. 69, 414-422.

[17] Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment, *British Journal of Radiology,* Vol. 44, 793-797.

[18] Jennison, C. (1987). Efficient group sequential tests with unpredictable group sizes, *Biometrika,* Vol. 74, 155-165.

[19] Jennison, C., and Turnbull, B. W. (1983). Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials, *Techometrics,* Vol. 25, 49-58.

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

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

[22] Kim, K., and DeMets, D. L. (1987-a). Design and analysis of group sequential tests based on the type I error spending rate function, *Biometrika,* Vol. 74, 149-154.

[23] Kim, K., and DeMets, D. L. (1987-b). Confidence intervals following group sequential tests in clinical trials, *Biometrics,* Vol. 43, 857-864.

[24] Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics,* Vol. 38, 963-974.

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

[26] Lan, K. K. G., Simon, R., and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials, *Communications in Statistics,* C (Sequential Analysis). Vol. 1, 207-219.

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

[28] Lee, J. W., and DeMets, D. L. (1991). Sequential comparison of changes with repeated measurements data, *Journal of the American Statistical Association,* Vol. 86, 757-762.

[29] Lee, J. W., and DeMets, D. L. (1992). Sequential rank tests with repeated measurements in clinical trials, *Journal of the American Statistical Association,* Vol. 87, 136-142.

[30] Lee, J. W., and DeMets, D. L. (in press). Group sequential comparison of changes: ad-hoc vs. more exact method, *Biometrics.*

[31] Lee, J. W., Reboussin, D. M., and DeMets, D. L. (1990). Rank tests for multivariate linear models in the presence of missing data, Technical Report 59, University of Wisconsin-Madison, Biostatistics Center.

[32] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized

linear models, *Biometrika,* Vol. 73, 13-22.

[33] Lindstrom, M. J., and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data, *Journal of the American Statistical Association,* Vol. 83, 1014-1022.

[34] 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, Vol. 134, 15-25.

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

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

[37] Pocock, S. J. (1982). Interim analyses for randomized clinical trials: The group sequential approach, *Biometrics,* Vol. 38, 153-162.

[38] 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,* Vol. 48, 1131-1143.

[39] Puri, M. L., and Sen, P. K. (1971). *Nonparametric methods in multivariate analysis,* John Wiley and Sons, New York.

[40] Puri, M. L., and Sen, P. K. (1985). *Nonparametric methods in general linear models,* John Wiley and Sons, New York.

[41] 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,* Vol. 77, 862-868.

[42] Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials, *Statistics in Medicine,* Vol. 5, 421-433.

[43] Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power?, *Controlled Clinical Trials,* Vol. 7, 8-17.

[44] Su, J. Q., and Lachin, J. M. (1992). Group sequential distribution-free methods for the analysis of multivariate observations, *Biometrics,* Vol. 48, 1033-1042.

[45] Wald, A. (1947). *Sequential Analysis.* John Wiley and Sons, New York.

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

[47] Ware, J. H. (1985). Linear models for the analysis of longitudinal studies, *The American Statistician,* Vol. 39, 95-101.

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

[49] Wei, L. J., Su, J. Q., and Lachin, J. M. (1990). Interim analyses with repeated measurements in a sequential clinical trial, *Biometrika,* Vol. 77, 359-364.

[50] Wu, M. C. and Lan, K. K. G. (1992). Sequential monitoring for comparison of changes

in a response variable in clinical studies, *Biometrics*, Vol. 48, 765-780.

[51] Wu, M. C. and Bailey, K. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model, *Biometrics*, Vol. 45, 939-955.

집단축차검정법들에 관한 고찰

이재원[2)]

요약

윤리적 또는 경제적 이유 때문에 임상실험(Clinical Trials)의 연구자들은 실험중간에 새로운 치료방법이나 약이 효과가 있는지 또는 해로운지를 반복해서 검정한다. 하지만 각 검정마다 정해진 유의수준을 반복해서 사용하면 전체 유의수준이 상당히 커지게 된다. 임상실험에서 발생하는 이러한 문제를 해결하기 위하여 많은 집단축차검정법 (Group Sequential Testing Procedure)들이 개발되어 왔다. 본 논문에서는 이러한 집단축차검정법들을 간략하게 비교분석하고자 한다.

2) (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과.