BIOMETRICS 55, 853–857
September 1999

Modification of Sample Size in Group Sequential Clinical Trials

Lu Cui,[1,*] H. M. James Hung,[1] and Sue-Jane Wang[2]

[1]Division of Biometrics I, OB/CDER, Food and Drug Administration,
1451 Rockville Pike, Rockville, Maryland 20852, U.S.A.
[2]Division of Biometrics II, OB/CDER, Food and Drug Administration,
5600 Fishers Lane, Rockville, Maryland 20857, U.S.A.
[*] *email:* cuil@cder.fda.gov

SUMMARY. In group sequential clinical trials, sample size reestimation can be a complicated issue when it allows for change of sample size to be influenced by an observed sample path. Our simulation studies show that increasing sample size based on an interim estimate of the treatment difference can substantially inflate the probability of type I error in most practical situations. A new group sequential test procedure is developed by modifying the weights used in the traditional repeated significance two-sample mean test. The new test has the type I error probability preserved at the target level and can provide a substantial gain in power with the increase of sample size. Generalization of the new procedure is discussed.

KEY WORDS: Interim analysis; Power; Sample size; Type I error.

1. Introduction

Determination of sample size rests upon knowledge of the expected treatment effect size, which is a function of the expected treatment difference and the variance of an outcome variable. If the actual variance is much larger than expected, the planned sample size will be severely underestimated and consequently it may not be sufficient to give reasonable power to show the treatment efficacy. For nonsequential trials, several authors have studied ways of reestimating sample size based on reevaluation of the nuisance parameters at some interim stage of a trial (Wittes and Brittain, 1990; Gould, 1992; Gould and Shih, 1992; Shih, 1992, 1993).

A variety of group sequential methods (e.g., Pocock, 1977, 1982; O'Brien and Fleming, 1979; Lan and DeMets, 1983) have been used to allow early termination of the trials due to overwhelming efficacy or undue harm. With interim analyses, treatment codes are often broken in order to perform repeated significance tests. Consequently, when the observed sample path indicates that the estimated treatment difference is greatly different from the expectation, one would attempt to modify the planned sample size or the so-called maximum information (termed by Lan and DeMets, 1983). Recently, Proschan and Hunsberger (1995) proposed a method of extending a nonsequential study based on the observed treatment difference. In the sequential setting, Lan and Trost (1997) explored the possibility of modification of sample size based on the observed treatment difference.

The research work of this paper is motivated by many case studies from the new drug applications or study protocols reviewed by the Food and Drug Administration during the past few years, in which sample size increase based on the observed sample path was proposed during the midcourse of a group sequential trial. For example, in one case study involving a Phase III group sequential clinical trial for evaluating the effect of a new drug for prevention of myocardial infarction in patients undergoing coronary artery bypass graft surgery, the sample size of 600 subjects per treatment group was originally planned to detect (two-sample proportion test) a 50% reduction of incidence, or a change in the incidence from 22% for placebo to 11% for the drug, with 95% power. However, based on the interim analysis on the data from 50% of the planned population, the probability of finding an ultimate, statistically significant result was very low if the present trend continued for the remaining duration of the study. The incidence rate in the placebo group was in line with the expectation, but the observed incidence for the drug was about 16.5%, about a 25% reduction, and far below the 50% target originally assumed. A proposal to expand sample size was therefore submitted. At the time, the general feeling was that such an increase of sample size may substantially inflate the overall type I error rate. No valid testing procedure was available in literature to account for such an outcome-dependent adjustment of sample size. Partly because of this reason, the drug sponsor decided not to increase sample size and the trial eventually failed to show a statistically significant treatment effect. This real case study reflects a major difficulty in planning clinical trials, i.e., the need to obtain an accurate estimate of the treatment effect for planning the trial with little prior knowledge of the effect. A natural solution to this dilemma is implementation of a valid inferential procedure that allows flexibility for adjusting sample size based on the updated estimate of treatment effect during the course of the trial.

2. Impact of Sample Size Change Based on an Interim Estimate of Treatment Difference

Consider, for simplicity, the problem of detecting a difference in the means of the two independent normal populations with a known variance σ^2, say $\sigma^2 = 1$. Let μ_1 and μ_2 be the respective means and let $\Delta = \mu_1 - \mu_2$. Of interest is to test the null hypothesis H: $\Delta = 0$ versus the alternative hypothesis A: $\Delta > 0$ using the two-sample mean test. Let N denote the total sample size per population planned for detecting an expected treatment difference $\Delta = \delta$ at the desired significance level α and with power $1 - \beta$. Thus,

$$N = 2\{(z_\alpha + z_\beta)/\delta\}^2, \tag{2.1}$$

where z_α is the $(1-\alpha)$th percentile of the standard normal distribution.

Suppose that it is planned to perform up to $K-1$ interim analyses and one possible final analysis and that n_k subjects are obtained for each population between the $(k-1)$th and kth analyses. Let N_k be the total number of subjects obtained up to the kth analysis for each population and let $t_k = N_k/N$ be the information fraction or information time at the kth interim analysis. Denote by T_k the two-sample mean test statistic at time t_k. An α spending function $\alpha(t)$ is selected prior to the start of the trial, where t indexes the information time, $\alpha(0) = 0$, and $\alpha(1) = \alpha$. The critical values C_k are then determined (Lan and DeMets, 1983).

At the completion of the Lth interim analysis for some specified L $(1 \le L \le K-1)$, the observed treatment difference Δ_L deviates to a large extent from the expected difference δ so that a decision to adjust the sample size is prompted. The sample size may be adjusted based strictly on the difference between Δ_L and δ or based on conditional power (Lan, Simon, and Halperin, 1982). In the group sequential design setting, the conditional power evaluated at the Lth interim analysis can be defined by

$$\begin{aligned} CP_L(\Delta) &= \mathrm{pr}(T_K > C_K \mid T_L, \Delta) \\ &= \Phi\Big(-\Big\{C_K - T_L t_L^{1/2} - (1-t_L)(N/2)^{1/2}\Delta\Big\} \\ &\qquad \div (1-t_L)^{1/2}\Big), \end{aligned}$$

where $\{T_K > C_K\}$ gives the rejection region for H at the final analysis. With the conditional power, one may adjust the sample size according to the following plan:

If $CP_L(\Delta_L) < \gamma_I CP_L(\delta)$ for some given positive constant $\gamma_I \le 1$ or $CP_L(\Delta_L) > \gamma_D CP_L(\delta)$ for some given $\gamma_D \ge 1$ at the Lth interim look, then adjust (increase or decrease) the total per group sample size by

$$M = N(\delta/\Delta_L)^2. \tag{2.2}$$

In theory, if sample size can be increased without bound based on the observed treatment difference, the overall type I error rate will be inflated greatly. In practice, there is often an upper limit for sample size, i.e., $M \le N_{\max}$ for some $N_{\max} > N$. It is worth noting that criterion (2.2) preserves the unconditional power at $1-\beta$ when $\Delta = \Delta_L$ and it gives a large M only when the observed treatment difference Δ_L is small.

Simulation studies are conducted to assess the impact of sample size modification based on the above plan on type I error rate and power. We set the total sample size per group to $N = 250$ for detecting $\delta = 0.30$ at $\alpha = 0.025$ and with power 0.90. Four interim analyses and one final analysis are to be performed at equal time intervals, i.e., the corresponding information times for the analyses are $k/5$, $k = 1, 2, 3, 4, 5$. The O'Brien–Fleming-type α spending function is chosen and the critical values for the repeated significance two-sample mean tests are computed using the program of DeMets, Kim, Lan, and Reboussin (available at the web site http://www.medsch.wisc.edu/landemets/). We set the upper limit of the adjusted total sample size to $N_{\max} = 4N$, $\gamma_I = 0.8$, and $\gamma_D = 1$. For the power study, we set the true treatment difference $\Delta = 0.21$, which gives a power of 60% with 250 subjects per group. The number of replications is set to 1,000,000 for evaluation of type I error rate and 20,000 for power unless stated otherwise. Suppose that we allow adjustment of sample size based on an interim estimate of Δ. Table A1 reveals that increasing sample size based on the observed treatment difference can improve power greatly (from 60% to 90% in our design setting). The large gain in power increase is at the cost of a substantial inflation in type I error rate (Table A1). Our study suggests that decreasing sample size has a mild effect on type I error rate and power.

The work of Lan and Trost (1997) pointed out that, as a result of sample size adjustment, the α spending function also needs to be adjusted and so are critical values. Our simulation studies show that, with proper adjustment of the α spending function, the inflation of type I error rate due to sample size increase is only slightly smaller than that without adjustment.

3. A New Group Sequential Test Procedure

We consider, without loss of generality, sample size increase based on the interim estimation of treatment difference. As a result of such a sample size adjustment, the repeated significance two-sample mean test is no longer a Brownian motion process. In addition, the total sample size becomes a random variable depending on the treatment difference observed during the course of the trial.

Suppose that, at the Lth interim analysis, a decision is made regarding whether to increase sample size based on the observed treatment difference Δ_L. Let M be the resulting total per-group sample size and M_{L+j} be the sample size at the $(L+j)$th look. M and M_{L+j} are functions of the observed treatment difference Δ_L. If Δ_L indicates that no sample size increase is needed, then $M = N$ and $M_{L+j} = N_{L+j}$; otherwise, M is adjusted, say, according to (2.2) and so is M_{L+j}. Furthermore, assume

$$M_{L+j} = b(N_{L+j} - N_L) + N_L, \tag{3.1}$$

where $b = (M - N_L)/(N - N_L)$, $j = 1, \ldots, K-L$.

Without allowing sample size adjustment, the two-sample mean test used at the $(L+j)$th interim analysis can be written as

$$\begin{aligned} T_{L+j} = {} & T_L(N_L/N_{L+j})^{1/2} \\ & + W_{L+j}\{(N_{L+j} - N_L)/N_{L+j}\}^{1/2}, \end{aligned} \tag{3.2}$$

where

$$W_{L+j} = \frac{\sum_{i=N_L+1}^{N_{L+j}} (X_i - Y_i)}{\sqrt{2(N_{L+j} - N_L)}}$$

Why sample size adjustments inflate type I error rates?

and X_i and Y_i are the ith observations in the two groups, respectively. When sample size is allowed to increase from N to M, this test becomes

$$T^*_{L+j} = T_L(N_L/M_{L+j})^{1/2} + W^*_{L+j}\{(M_{L+j} - N_L)/M_{L+j}\}^{1/2},$$

where

$$W^*_{L+j} = \frac{\sum_{i=N_L+1}^{M_{L+j}} (X_i - Y_i)}{\sqrt{2(M_{L+j} - N_L)}}.$$

Apparently, allowing the increase in sample size with the outcome Δ_L not only replaces the original standardized test statistic W_{L+j} of a fixed amount of information by a new standardized statistic W^*_{L+j} of a random amount of information for the data between the interim looks L and $(L+j)$ but also alters the weights associated with T_L and W_{L+j}. The new weights are random because M_{L+j} is a function of Δ_L. While updating W_{L+j} with W^*_{L+j} is necessary as more information is accumulated, consideration of having the weights not influenced by Δ_L seems natural.

We now construct a new test procedure directly from the original repeated significance two-sample mean test. By keeping the original weights in (3.2), updating W_{L+j} with W^*_{L+j} will lead to

$$\begin{aligned} U_{L+j} &= T_L(N_L/N_{L+j})^{1/2} \\ &\quad + W^*_{L+j}\{(N_{L+j} - N_L)/N_{L+j}\}^{1/2}. \end{aligned} \tag{3.3}$$

With the new test procedure, at the kth look $(k = 1, \ldots, L)$, reject H if and only if $T_k > C_k$; otherwise, continue the trial. At the Lth interim look, if $T_L \le C_L$, then determine whether to increase sample size and then continue the trial. At the $(L+j)$th interim look, $j = 1, 2, \ldots, K-L$, reject H if $U_{L+j} > C_{L+j}$. Note that, if no sample size increase is made at the Lth look, U_{L+j} reduces to the original test statistic T_{L+j}. Under H, we have

$$\mathrm{E}(U_{L+j} \mid T_L) = \mathrm{E}(T_{L+j} \mid T_L) = T_L(N_L/N_{L+j})^{1/2} \tag{3.4}$$

$$\mathrm{var}(U_{L+j} \mid T_L) = \mathrm{var}(T_{L+j} \mid T_L) = (N_{L+j} - N_L)/N_{L+j}, \tag{3.5}$$

and it follows from (3.1) that

$$\begin{aligned} &\mathrm{cov}(U_{L+i}, U_{L+j} \mid T_L) \\ &\quad = \{(N_{L+i} - N_L)(N_{L+j} - N_L)/(N_{L+i}N_{L+j})\}^{1/2} \\ &\qquad \times \{(M_{L+\min(i,j)} - N_L)/(M_{L+\max(i,j)} - N_L)\}^{1/2} \\ &\quad = \{(N_{L+i} - N_L)(N_{L+j} - N_L)/(N_{L+i}N_{L+j})\}^{1/2} \\ &\qquad \times \{(N_{L+\min(i,j)} - N_L)/(N_{L+\max(i,j)} - N_L)\}^{1/2} \\ &\quad = \mathrm{cov}(T_{L+i}, T_{L+j} \mid T_L), \end{aligned} \tag{3.6}$$

where $1 \le i, j \le K - L$ and $i \ne j$. Under H, conditional on T_L, the vectors $(U_{L+1}, \ldots, U_K)$ and $(T_{L+1}, \ldots, T_K)$ have the same multivariate normal distribution. Hence, the joint distribution of $(T_1, \ldots, T_L, U_{L+1}, \ldots, U_K)$ is the same as that of $(T_1, \ldots, T_L, T_{L+1}, \ldots, T_K)$, leading to

$$\begin{aligned} &\mathrm{pr_H}\left(\left\{\bigcup_{k=1}^{L}(T_k > C_k)\right\} \cup \left\{\bigcup_{k=L+1}^{K}(U_k > C_k)\right\}\right) \\ &\quad = \mathrm{pr_H}\left(\bigcup_{k=1}^{K}(T_k > C_k)\right). \end{aligned} \tag{3.7}$$

Thus, the total type I error probability of the new test procedure allowing sample size increase is equal to that of the original test procedure without increase, which is equal to α.

Now consider the power of the new test when a sample size increase is allowed. For $j = 1, \ldots, K - L$, let $d_{L+j}(T_L) = \mathrm{E}(U_{L+j} \mid T_L) - \mathrm{E}(T_{L+j} \mid T_L)$. If the sample size is increased after the Lth interim analysis, under the alternative hypothesis A,

$$\begin{aligned} d_{L+j}(T_L) &= [T_L(N_L/N_{L+j})^{1/2} + \{(M_{L+j} - N_L)/2\}^{1/2} \\ &\qquad \times \{(N_{L+j} - N_L)/N_{L+j}\}^{1/2}\Delta] \\ &\quad - [T_L(N_L/N_{L+j})^{1/2} + \{(N_{L+j} - N_L)/2\}^{1/2} \\ &\qquad \times \{(N_{L+j} - N_L)/N_{L+j}\}^{1/2}\Delta] > 0, \\ &\qquad 1 \le j \le K - L. \end{aligned}$$

Since, under the alternative, the variance–covariance structure given by (3.5) and (3.6) still holds, the joint distributions of $(T_1, \ldots, T_L, U_{L+1}, \ldots, U_K)$ and $(T_1, \ldots, T_L, T_{L+1} + d_{L+1}(T_L), \ldots, T_K + d_K(T_L))$ are identical. Thus, under the alternative hypothesis A, it follows immediately that the lefthand side of (3.7) is larger than the righthand side. Hence, the power is greater with the new test procedure.

Monte Carlo simulation was conducted to estimate the size and the gain in power for the new test. Table A2 gives the empirical type I error rate and power of the new test procedure for the Gaussian outcome in the cases of equal and unequal time intervals. Clearly, the new test has its type I error rate maintained at the 0.025 level and the power can increase from 60%, the power level for no sample size increase, to 93%.

4. Generalization

By use of the idea for the construction of the new group sequential two-sample mean test given above, we can develop a new group sequential test based on the repeated significance test that can be asymptotically expressed as a Brownian motion process, e.g., logrank test, with independent increment property (Lan and DeMets, 1983; Lan and Wittes, 1988). Let $B(t)$ be such a repeated significance test evaluated at the information time t with variance t. In this test procedure, the maximum information is scaled to one and $0 \le t \le 1$. Let $T(t) = B(t)/t^{1/2}$, the standardized version of $B(t)$. Suppose that the decision to increase the maximum information from one to ω is made at time $t = t_L$ on the basis of the observed value of $T(t_L)$. Let $b = (\omega - t_L)/(1 - t_L)$. Then $b \ge 1$. As before, the new sequential test procedure resulting from $T(t)$ can be constructed as

$$U(t) = T(t) \quad \text{if } t \le t_L,$$

and

$$\begin{aligned} U(t) &= T(t_L)\{w(t_L, t)\}^{1/2} \\ &\quad + [\{B(b(t - t_L) + t_L) - B(t_L)\}/\{b(t - t_L)\}^{1/2}] \\ &\quad \times [1 - w(t_L, t)]^{1/2}, \qquad t_L \le t \le 1, \end{aligned}$$

where $w(t_L, t) = t_L/t$. By arguments similar to those given in Section 3, under the null hypothesis, the new test $U(t)$ and the original test $T(t)$ have the same finite element distributions. Therefore, using the original rejection boundary for $T(t)$, the new test $U(t)$ will have the total type I error probability preserved at the specified level. It can have substantial gain in power with an increase of maximum information.

In the example of prevention of myocardial infarction discussed earlier, if the incidence rates observed at the interim look are the true rates for the two treatment groups, the power of the test is only about 40%. Our simulation shows that, if the per-group sample size were increased to about 1400, using the new group sequential two-sample proportion test $U(t)$, the power would be increased to 93%, doubling the probability for the trial to conclude the drug efficacy. The type I error rate associated with the sample size adjustment is 0.024. In the simulation, a Bernoulli outcome was generated and 20,000 replications were used.

5. Discussion

The new group sequential test procedure uses the same critical values as those for the original test procedure. The type I error probability is also preserved when the sample size is allowed to decrease based on the interim estimate of treatment difference. From the mathematical arguments given above, the new group sequential test remains valid when the time for consideration of sample size adjustment is not prespecified prior to the first interim analysis as long as it does not depend on the future data.

Acknowledgements

The authors wish to thank the referee and the associate editor for comments that helped improve this paper. This research was supported by Review Science Research Fund RSR-096-010A, Center for Drug Evaluation and Research, U.S. Food and Drug Administration. The opinions expressed herein are those of the authors and are not necessarily those of the U.S. Food and Drug Administration.

Résumé

Dans les essais cliniques séquentiels groupés, la ré-estimation de la taille de l'échantillon peut être un point compliqué quand il est permis au changement d'effectif d'être influencé par le chemin observé. Nos études de simulation montrent qu'accroître la taille de l'échantillon basé sur une estimation intermédiaire de la différence liée au traitement peut substantiellement accroître la probabilité d'une erreur de type I dans la plupart des situations pratiques. Une nouvelle procédure de test séquentiel groupé est développée en modifiant les poids utilisés dans le test traditionnel de signification pour la comparaison répétée de deux moyennes. Le nouveau test à une probabilité d'erreur de type I préservée et peut fournir des gains substantiels en puissance avec l'accroissement de la taille de l'échantillon. Une généralisation de la nouvelle procédure est discutée.

References

Gould, A. L. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine* **11,** 55–66.

Gould, A. L. and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics—Theory and Methods* **21**(10), 2833–2853.

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

Lan, K. K. G. and Trost, D. C. (1997). Estimation of parameters and sample size reestimation. Proceedings of Biopharmaceutical Section, American Statistical Association.

Lan, K. K. G. and Wittes, J. (1988). The B-value: A tool for monitoring data. *Biometrics* **44,** 579–585.

Lan, K. K. G., Simon, R., and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics* **C1,** 207–219.

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

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

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

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51,** 1315–1324.

Shih, W. J. (1992). Sample size reestimation in clinical trials. In *Biopharmaceutical Sequential Statistical Applications,* K. Peace (ed), 285–301. New York: Marcel Dekker.

Shih, W. J. (1993). Sample size reestimation for triple blind clinical trials. *Drug Information Journal* **27,** 761–764.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9,** 65–72.

Received June 1998. *Revised January* 1999.
Accepted February 1999.

Appendix

Table A1

Impact of sample size adjustment on power and type I error rate

Time of sample size change, t_L	Decreasing sample size only	Increasing sample size	Decreasing or increasing sample size
Power			
0.20	0.59	0.87	0.84
0.40	0.60	0.91	0.91
0.60	0.61	0.94	0.94
0.80	0.61	0.96	0.96
No sample size change	0.61	0.61	0.61
Type I error rate			
0.20	0.032	0.032	0.038
0.40	0.026	0.033	0.035
0.60	0.025	0.037	0.037
0.80	0.024	0.033	0.033
No sample size change	0.025	0.025	0.025

Table A2

Type I error rate and power of the new test increasing sample size only

Time of sample size change, t_L	Type I error rate	Power
Equal sample size increments		
0.20	0.025	0.86
0.40	0.025	0.90
0.60	0.025	0.92
0.80	0.025	0.91
No sample size change	0.025	0.61
Unequal sample size increments		
0.20	0.025	0.86
0.50	0.025	0.93
0.60	0.025	0.92
0.85	0.025	0.90
No sample size change	0.025	0.61