BIOMETRICS 55, 1286–1290
December 1999

Adaptive Sample Size Calculations in Group Sequential Trials

Walter Lehmacher* **and Gernot Wassmer**
Institut für Medizinische Statistik, Informatik und Epidemiologie,
Universität zu Köln, Joseph-Stelzmann-Str. 9, D-50931 Köln, Germany
* *email:* walter.lehmacher@medizin.uni-koeln.de

SUMMARY. A method for group sequential trials that is based on the inverse normal method for combining the results of the separate stages is proposed. Without exaggerating the Type I error rate, this method enables data-driven sample size reassessments during the course of the study. It uses the stopping boundaries of the classical group sequential tests. Furthermore, exact test procedures may be derived for a wide range of applications. The procedure is compared with the classical designs in terms of power and expected sample size.

KEY WORDS: Adaptive interim analysis; Clinical trials; Group sequential tests; Sample size.

1. Introduction

The classical group sequential test procedures that are due to Pocock (1977) and O'Brien and Fleming (1979) assume the number of the sequentially planned stages of the clinical trial to be fixed in advance and the number of observations to be equal between the stages. It is well known that small deviations from the equal sample size allocation between the stages do not have a great influence on the Type I error rate and can, in most cases, be ignored. The use function approach (Lan and DeMets, 1983), on the other hand, explicitly incorporates unequal sample sizes between the analyses and the number of the stages must not be fixed. In particular, this method can be applied if the group sizes in the subsequent stages are not known in advance. The most common application is given if the interim analyses are performed at specific but arbitrary time points rather than after a specified number of observations.

The distinctive feature of the classical methods is that they require a data-independent choice of the group sizes. Only external factors may influence the modification of the design in the ongoing trial. It is tempting, however, to take into account current data trends. In particular, it should be possible to adjust or reassess sample size calculations that are merely based on knowledge available prior to the study. Such an adaptive planning is indicated if, e.g., the sample variance is higher than anticipated. Then it is natural to plan more observations than originally planned to reach a desired power that is conditioned on the observed effect size. Stein (1945) has already considered this. He proposed a two-stage procedure where the variance of the first stage is used for planning the second stage. The observed effect size, however, may not be taken into account. Essentially, there are two methods that are specifically designed for this purpose. The first (Bauer and Köhne, 1994) is based on Fisher's combination test and the second (Proschan and Hunsberger, 1995) is based on the specification of a conditional error function. Both are mainly designed for only one interim analysis and testing against a one-sided hypothesis. It is shown by Wassmer (1998) that, for these cases, the methods are virtually identical in terms of power and average sample size.

In the present paper, we propose a strategy that uses the classical stopping boundaries while enabling an adaptive planning of the ongoing trial. The idea is very simple. If, at some stage k, one always uses the unweighted mean of the test statistics and the critical values designed for the case of equal sample sizes between the stages, then the resulting group sequential test procedure is independent of these sample sizes as long as the test statistics are independent and standard normally distributed. These normal scores are obtained by the inverse normal method, which is a common method in combining test results. Applying this method in a group sequential setting offers a wide range of applications, as will be shown.

In the next section, the proposed strategy will be described in detail. It will be shown that the classical group sequential designing schemes can be used within the application of the proposed adaptive procedure. We compare the power and the expected sample size of our procedure with the classical nonadaptive procedures. A pharmaceutical study example is given. In the Discussion, we briefly outline some general aspects for the design of an adaptively planned clinical trial.

2. The Proposed Procedure

2.1 *Classical Group Sequential Test Procedures*

The basic concept of the classical group sequential test designs for the two-sample comparisons in clinical trials is described as follows. Consider K groups (stages) of normally distributed observations, where in group k, $k = 1, 2, \ldots, K$, n_k observations are observed in sample $i, i = 1, 2$. Let $\bar{X}_{ik}$ denote the mean response of sample i in the kth group of observations. Furthermore, let the variance σ^2 be known and constant over the stages and the two subsamples. In the kth stage, the normal score Z_k is given by

$$Z_k = \frac{\bar{X}_{1k} - \bar{X}_{2k}}{\sigma}\sqrt{\frac{n_k}{2}}. \tag{1}$$

The cumulative standardized normal score

$$Z_k^* = \left(\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}\right)^{-1/2} \sum_{\tilde{k}=1}^{k} \sqrt{n_{\tilde{k}}}\, Z_{\tilde{k}} \tag{2}$$

is the usual statistic for testing the hypothesis $\text{H}_0\colon \mu_1 - \mu_2 = 0$, where μ_1 and μ_2 denote the unknown means in the two subsamples that combine all relevant information available up to some stage k. Both Z_k and Z_k^* are standard normally distributed under H_0.

Consider the case of equal sample size allocation between the stages (i.e., $n_1 = n_2 = \cdots = n_K = n$), where the test statistic (2) reduces to $\Sigma_{\tilde{k}=1}^{k} Z_{\tilde{k}}/k^{1/2}$. For testing the null hypothesis H_0 against the two-sided alternative $\text{H}_1\colon \mu_1 - \mu_2 \neq 0$, Wang and Tsiatis (1987) proposed the Δ class of tests with critical values $u_k = k^{\Delta - 0.5}\, c(\alpha, K, \Delta)$, $k = 1, 2, \ldots, K$. For given K and Δ, one has to find the critical values u_k such that, under H_0,

$$P(|\, Z_1^* \,| < u_1, |\, Z_2^* \,| < u_2, \ldots, |\, Z_K^* \,| < u_K) = 1 - \alpha, \tag{3}$$

which ensures that the test procedure controls the level α. Specifically, setting $\Delta = 0$, one gets O'Brien and Fleming's (1979) monotonically decreasing critical values, and setting $\Delta = 0.5$, one gets Pocock's (1977) procedure with constant critical values at each stage k of the group sequential test procedure.

In the classical concept, these critical values are, in an approximative sense, also used for unequal sample size allocations between the stages where the test statistic is given by (2). Furthermore, for the more realistic case of an unknown variance σ^2, the pooled sample variance of observations up to some stage k is used as an estimate for σ^2 and is used for σ^2 in (2). Both are known to not substantially affect the Type I error rate α for many cases (Pocock, 1982). An alternative exact solution that additionally provides the possibility of adaptively choosing the sample size based on the data observed so far is presented in the next section.

2.2 *Adaptive Group Sequential Test Procedures*

As in the last section, consider the case of a two-sample comparison of means with normally distributed observations and known σ^2; $n_1, n_2, \ldots, n_K$ are not assumed to be equal. If one uses

$$\frac{1}{\sqrt{k}} \sum_{\tilde{k}=1}^{k} Z_{\tilde{k}}, \tag{4}$$

where $Z_{\tilde{k}}$ is as given in (1), and the classical group sequential boundaries (derived for equal sample size allocation), then the procedure controls the level α for any sequence $n_1, n_2, \ldots, n_K$ since Z_k, $k = 1, 2, \ldots, K$, are (conditionally) independent and standard normally distributed, irrespective of how the n_k's are determined. All information available at some stage k can be used for the determination of the sample sizes for the subsequent stages. The proposed approach maintains α exactly for any possibly data-driven choice of sample sizes. The formal proof for this property is given in Bauer (1989).

The approach requires the use of the unweighted normal scores (4) also for unequal and adaptively chosen sample sizes. This might seem counterintuitive since this method does not explicitly account for different sample sizes. Nevertheless, it is impressive because of its practical applicability as the critical values of the classical approaches can be used. It is shown in the next section that it is hardly less powerful than existing methods for unequal sample sizes. Note that it is not even necessary to assume the number of observations in the subsamples to be equal in each analysis stage.

Using our approach, one may also consider a weighted sum of the Z scores where the weights are preset, e.g., to reflect likely patterns of information collection. A weighting scheme might be helpful if no equal sample sizes n_k between the K sequences are planned. This may be the case if, e.g., it is decided to perform the first look later than after n subjects for enlarging the power of the first stage trial or if the first stopping time is chosen earlier due to administrative reasons. In this case, the critical values are derived from (3) using the specific weighting scheme. Using the proposed approach, these preset values must be used throughout the trial. It should be pointed out that asymptotic unbiasedness is conserved when the weights in (2) are chosen adaptively, as long as the choices depend only on previously observed data through variance estimates and not on effect size estimates. This is shown by simulation findings in the work of Gould and coworkers (Gould and Shih (1998) and references therein). However, our method of adapting the sample sizes is exact and based on the complete unblinded data. It is not restricted to adaptive sample size reassessments that are merely based on variance estimates.

2.3 *Repeated Confidence Intervals*

Based on the adaptive testing procedure described above, it is easy to derive the corresponding repeated confidence intervals (RCIs) that are introduced by Jennison and Turnbull (1989). In the case of known σ^2, at each stage k, the two-sided $(1 - \alpha)100\%$ RCI is given by

$$\left(\sum_{\tilde{k}=1}^{k} \sqrt{n_{\tilde{k}}}\right)^{-1} \sum_{\tilde{k}=1}^{k} \sqrt{n_{\tilde{k}}}\,(\bar{X}_{1\tilde{k}} - \bar{X}_{2\tilde{k}}) \pm u_k \sigma \sqrt{k} \left(\sum_{\tilde{k}=1}^{k} \sqrt{\frac{n_{\tilde{k}}}{2}}\right)^{-1}, \tag{5}$$

where u_k denotes the critical value of a two-sided group sequential level-α test. The formula illustrates the weighting scheme for the sample means that is the basis for the proposed approach. Note that the RCIs are valid whatever stopping criterion is employed and hence are different from confidence intervals following a group sequential test.

2.4 *A General Method*

The test statistic that results from the inverse normal method of combining independent p values (Hedges and Olkin, 1985) is given by

$$\frac{1}{\sqrt{k}} \sum_{\tilde{k}=1}^{k} \Phi^{-1}\left(1 - p_{\tilde{k}}\right), \tag{6}$$

Table 1

Power and average sample number (asn) per treatment group of the proposed two-sided adaptive group sequential test procedure for specified sample size allocation vector W as compared to the original approach as if W was known. $K = 5$, $\alpha = 0.05$, $1 - \beta =$ power of equally spaced original approach with $W = (20, 40, 60, 80, 100)$; Pocock's design with constant critical values.

	$1-\beta=0.80$				$1-\beta=0.90$			
	Proposed		W known		Proposed		W known	
W	Power	asn	Power	asn	Power	asn	Power	asn
(10, 20, 30, 60, 100)	0.722	69.0	0.775	68.5	0.844	60.5	0.884	60.0
(10, 20, 30, 100, 160)	0.878	98.3	0.938	93.7	0.953	83.5	0.982	80.1
(10, 20, 100, 140, 200)	0.955	109.3	0.978	102.0	0.989	93.7	0.996	89.0
(20, 40, 60, 80, 100)	0.800	65.0	0.800	65.0	0.900	56.8	0.900	56.8
(20, 40, 80, 120, 160)	0.937	83.1	0.944	82.6	0.982	68.8	0.984	68.7
(20, 80, 120, 160, 200)	0.978	95.4	0.981	93.2	0.996	81.3	0.997	79.7
(40, 60, 80, 90, 100)	0.823	68.0	0.824	66.9	0.914	60.9	0.915	59.8
(40, 100, 140, 150, 160)	0.957	91.5	0.957	89.0	0.988	79.3	0.989	77.2
(40, 80, 100, 150, 200)	0.978	87.4	0.981	87.6	0.996	72.8	0.997	73.0

where $\Phi^{-1}(\cdot)$ denotes the inverse cumulative standard normal distribution function. The proposed approach involves using the classical group sequential boundaries for the test statistic (6). Since the $\Phi^{-1}(1-p_k)$'s, $k = 1, 2, \ldots, K$, are independent and standard normally distributed, the proposed approach maintains α exactly for any (adaptive) choice of sample sizes.

There are several potential applications for using this method. In general, any testing situation that reaches exact p values for the independent stages of the group sequential design can be adopted. For example, for the two-sample comparison of normally distributed outcomes with unknown σ^2, consider the stochastically independent one-sided p values p_k, $k = 1, 2, \ldots, K$, of the separate t-test statistics based on n_k observations per sample. Then, using the test statistic (6) and the classical group sequential boundaries results in an exact level-α adaptive group sequential test procedure. Note that in (6) it is necessary to use the one-sided p values for the one-sided and the two-sided case to avoid directional conflicts, i.e., the case that opposite effects at previous stages may lead to the rejection of H_0 at some stage k of the procedure.

3. Power Assessments

In order to assess the power performance of the procedure, we treat the case of the two-sample comparison with known variance. The proposed procedure is compared with the classical group sequential test designs with given and fixed sample sizes n_k, $k = 1, 2, \ldots, K$, i.e., the critical values of the classical group sequential test were calculated as if the specific sequence of sample sizes $n_1, n_2, \ldots, n_K$ was known, and the corresponding power and average sample sizes for a specified alternative were calculated. This approach is in some sense optimal, as it corresponds to using the uniformly most powerful unbiased test at each stage of the group sequential design. Actually, the comparison with the proposed approach is artificial since it does not account for the adaptive nature of our design. On the other hand, it is a fair comparison of exact level-α test procedures, which might also provide a rough evaluation of the power loss that is due to adaptively planning the sample sizes in the ongoing trial. In Table 1, we provide the power values and average sample numbers (asn) for $K = 5$ in Pocock's design. The critical values for a specified sample size allocation $W = (n_1, n_1 + n_2, \ldots, n_1 + n_2 + \cdots + n_K)$ and the test characteristics were calculated using SAS/IML software (SAS Institute, 1995) as described in Wassmer (1999). The effect size was chosen such that the approaches reach power 0.80 and 0.90 for the case of an equal sample size allocation with $n_1 = n_2 = \cdots = n_5 = 20$.

Table 1 describes the typical power performance of the proposed approach. Essentially, the loss in power is surprisingly small except for the case of low power in the earlier stages of the study ($n_1 = 10$). In the case with low power in the early stages, the loss is increasing for increasing numbers of early looks. This only happens, however, for sample sizes that rarely occur in clinical practice. The loss in power that is due to adaptively designing the study is comparable to that described earlier for adaptive designs (Banik, Köhne, and Bauer, 1996; Bauer and Köhne, 1994; Proschan, Follmann, and Waclawiw, 1992).

4. Example

In a randomized, placebo-controlled, double-blind study involving patients with acne papulopustulosa, Plewig's grade II–III, the effect of treatment under a combination of 1% chloramphenicol (CAS 56-75-7) and 0.5% pale sulfonated shale oil versus the alcoholic vehicle (placebo) was investigated (Fluhr et al., 1998). After 6 weeks of treatment, reduction of bacteria from baseline, examined on agar plates (log CFU/cm^2; CFU, colony forming units), of the active group as compared to the placebo group were assessed. The available data were from 24 and 26 patients in the combination drug and the placebo groups, respectively. The combination therapy resulted in a highly significant reduction in bacteria as compared to placebo using a two-sided t-test for the changes ($p = 0.0008$).

To illustrate the general method proposed in this paper, suppose that it was intended to perform a three-stage adaptive Pocock's design with $\alpha = 0.01$ and after 2 × 12 patients the first interim analysis was planned. The two-sided critical bounds for this method are $u_1 = u_2 = u_3 = 2.873$ (Pocock, 1977). After $n_1 = 12$ patients per group, the test statistic of the t-test is 2.672 with one-sided p value $p_1 = 0.0070$, resulting from an observed effect size $\bar{x}_{11} - \bar{x}_{21} = 1.549$ and an observed standard deviation $s_1 = 1.316$. The study should be continued since $\Phi^{-1}(1 - p_1) = 2.460 < u_1$. The observed effect is fairly near to significance. We therefore plan the second interim analysis to be conducted after observing the next 2×6 patients, i.e., the second interim analysis will be performed after fewer patients than the first. The t-test statistic of the second stage is equal to 1.853 with one-sided p value $p_2 = 0.0468$ ($\bar{x}_{12} - \bar{x}_{22} = 1.580$, standard deviation of the second stage $s_2 = 1.472$). The test statistic (6) becomes $2^{-1/2}(\Phi^{-1}(1 - p_1) + \Phi^{-1}(1 - p_2)) = 2.925$, yielding a significant result after the second stage of the trial. Corresponding approximate 99% RCIs may be found by replacing σ in (5) with the observed standard deviation, which yields the intervals $(-0.12, 3.21)$ and $(0.21, 2.92)$ for the first and the second stages, respectively.

In this example, the application of the classical Pocock procedure (i.e., the approximate use of the critical values for the usual t-test statistics) would lead to the same statistical result since the global t-test statistic from the two stages also turns out to be greater than u_2. Note, however, that this is an incorrect application of the procedure since the sequential design is altered on the basis of the observed responses. On the other hand, the use of the proposed general method yields exact test results that are unbiased with respect to data-driven adaptations of the design.

5. Discussion

We propose a method that enables the adaptive planning of sample sizes during the course of a group sequential clinical trial. The principal advantage of the approach is that the previously collected data can influence the decision of how much data to take before the next interim analysis through other than the observed variability. It is based on the inverse normal method of combining independent tests and thereby provides an exact solution to the problem of unknown variance and unequal sample sizes between the stages of the trial. A wide range of designing variants developed in the theory of group sequential tests can be applied. These include one-sided group sequential test procedures that account for early stopping in favor of H_0 (DeMets and Ware, 1980, 1982). A disadvantage arises from the fact that K must be fixed prior to analyses. The use function approach does not assume this, but requires the subsequent sample sizes (or monitoring times) to be determined independently of current data trends (Proschan et al., 1992). In our approach, a data-dependent determination of the sample sizes is possible where also unbalanced sample sizes between the two subsamples may be fixed in an interim analysis.

In planning such an adaptive group sequential trial, the maximum number K of sequences, the choice of the sequential boundaries (i.e., one-sided or two-sided and the desired α spending), and the weighting scheme for the Z scores have to be chosen. In practical applications, a group sequential design can be planned taking into account all well-known advantages and disadvantages concerning the choice of these issues. One has the full flexibility in the planning stage, but all these possible options have to be fixed in the study protocol. They cannot be changed during the course of the trial. On the other hand, after each interim analysis, the sample sizes in the subsequent sequences can be re-estimated to any extent using all information from the data, not only the variance estimation. One has the additional flexibility of a data-driven recalculation of the sample sizes between the stages.

There are several methods for the determination of the sample sizes between the stages of the procedure. Conditional power calculations may be performed (Proschan and Hunsberger, 1995) that are based on the observed effect size at some stage k of the procedure. Furthermore, we refer to a method proposed by Gould and Shih (1998) for estimating the variability before unblinding of the data that is based on an EM algorithm.

The proposed method is based on the unweighted combination of the normal scores of the separate stages. In the classical group sequential context, this refers to the use of the unweighted test statistic also for the case of unequal sample sizes between the stages. This is the difference between our method and the classical approach. Nevertheless, our method coincides with the classical approach if no adaption is necessary. If adaption is necessary, our method yields an exact adaptive level-α group sequential test procedure with no material loss in power as compared to the optimal test. A weighting scheme corresponding to planned nonequal sample sizes between the stages can be chosen. Corresponding repeated confidence intervals are easily derived. Hence, the proposed procedure offers a tool in group sequential analyses that copes well with the demands of practice.

An approach that is based on Fisher's method of combining p values, which was designed for up to two interim analyses, was proposed by Bauer and Köhne (1994). This approach seems to behave similarly to our approach. Both methods allow the adaptive planning of the sample sizes and may be applied to a wide range of situations due to the nonparametric use of the p values. To the contrary, the adaptive approach of Proschan and Hunsberger (1995) was specifically developed for normal responses with known variance and for one interim analysis. Further research will be necessary to generalize and assess these approaches. In particular, the consideration of the optimality aspect in the stagewise sample size calculations requires further investigations.

Acknowledgements

We would like to thank the editor and the associate editor for valuable comments that have improved the contents and the presentation of this paper.

Résumé

Une méthode pour essais séquentiels qui repose sur l'utilisation de l'inverse de la distribution normale pour combiner les résultats des étapes successives est proposée. Sans augmenter le risque d'erreur de type I, cette méthode permet de réévaluer l'effectif en cours d'étude en fonction des données disponibles. Elle fait appel aux même règles d'arrêt que les tests séquentiels classiques. En outre, des tests exacts peuvent être déduits pour un grand nombre d'applications. Cette procédure est comparée aux schémas classiques en termes de puissance et d'effectif moyen attendu.

References

Banik, N., Köhne, K., and Bauer, P. (1996). On the power of Fisher's combination test for two stage sampling in the presence of nuisance parameters. *Biometrical Journal* **38,** 25–37.

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie* **20,** 130–148.

Bauer, P. and Köhne, K. (1994). Evaluations of experiments with adaptive interim analyses. *Biometrics* **50,** 1029–1041. Correction in *Biometrics* **52,** 380.

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

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

Fluhr, J. W., Gloor, M., Merkel, W., Warnecke, J., Höffler, U., Lehmacher, W., and Glutsch, J. (1998). Antibacterial and sebosuppressive efficacy of a combination of chloramphenicol and pale sulfonated shale oil. *Arzneimittel-Forschung/Drug Research* **48**(I), 188–196.

Gould, A. L. and Shih, W. J. (1998). Modifying the design of ongoing trials without unblinding. *Statistics in Medicine* **17,** 89–100.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* New York: Academic Press.

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

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

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.

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

SAS Institute. (1995). *SAS/IML Software: Changes and Enhancements,* Release 6.11. Cary, NC: SAS Institute.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16,** 243–258.

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

Wassmer, G. (1998). A comparison of two methods for adaptive interim analysis in clinical trials. *Biometrics* **54,** 696–705.

Wassmer, G. (1999). Group sequential monitoring with arbitrary inspection times. *Biometrical Journal* **41,** 197–216.

Received June 1998.
Revised August and November 1998 *and February* 1999.
Accepted February 1999.