BIOMETRICS 57, 886–891
September 2001

Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical Group Sequential Approaches

Hans-Helge Müller* **and Helmut Schäfer**

Institute of Medical Biometry and Epidemiology, Philipps-University of Marburg,
Bunsenstraße 3, D-35037 Marburg, Germany
**email:* muellerh@mailer.uni-marburg.de or Hans-Helge.Mueller@med.uni-marburg.de

SUMMARY. A general method is presented integrating the concept of adaptive interim analyses into classical group sequential testing. This allows the researcher to represent every group sequential plan as an adaptive trial design and to make design changes during the course of the trial after every interim analysis in the same way as with adaptive designs. The concept of adaptive trial designing is thereby generalized to a large variety of possible sequential plans.

KEY WORDS: Adaptive design; Clinical trial; Combination rule; Group sequential test; Interim analysis.

1. Introduction

Interim analyses in controlled clinical trials are an important tool to reduce costs of the trial and risks for patients. They are performed for ethical and economic reasons. Two different statistical approaches have been proposed. The more classical statistical procedure is a group sequential statistical test, introduced and investigated by Pocock (1977, 1982), O'Brien and Fleming (1979), DeMets and Ware (1980, 1982), and Fleming, Harrington, and O'Brien (1984) and subsequently refined by many contributions such as Lan and DeMets (1983), Wang and Tsiatis (1987), Bauer (1992), Brittain and Bailey (1993), Müller and Schäfer (1999), and others. The approach of adaptive designs was introduced more recently by Bauer (1989) and Bauer and Köhne (1994) and, in a different way, by Proschan and Hunsberger (1995). Very recently, two relevant papers have been published by Cui, Huang, and Wang (1999) and by Lehmacher and Wassmer (1999). They presented methods allowing for a reassessment of the sample size after an interim analysis in a classical group sequential trial. Up to now, however, group sequential designs and adaptive designs have been viewed as alternative statistical approaches for planning and performing interim analyses in clinical trials. Each of these two approaches has obvious advantages compared with the other.

In group sequential designs, repeated significance testing is applied at predefined time points to cumulating data, with critical boundaries adjusted for multiple testing. In adaptive designs as proposed by Bauer and Köhne (1994), statistical tests are applied separately to the data of each stage of the trial and the resulting p-values are combined by combination rules such as Fisher's combination rule, making use of their stochastic independence. Alternatively, the independent test statistics from the different stages of the trial can be combined directly (Proschan and Hunsberger, 1995). These two types of adaptive procedures have been compared by Wassmer (1998). Power calculations for Fisher's combination test have been presented by Banik, Köhne, and Bauer (1996) and by Wassmer (1997).

The advantage of adaptive designs is their flexibility that allows the adaptation of certain experimental conditions, such as the sample size, the test statistic, or even the outcome variable used to measure the treatment effect, to the data observed in the previous stages of the trial. The advantage of group sequential tests is that there is a continuum of possible choices for the critical boundaries. In the planning phase, the group sequential test may optimally be fitted to the clinical situation and the special research problem. The researcher may choose a design out of a number of different plans published in the literature or may construct her or his own boundaries that meet the requirements of the individual trial, using fairly simple methods of numerical integration. For example, O'Brien and Fleming (1979) boundaries will lead to early stopping in case of unexpectedly large effects and will only slightly increase the maximum sample size over the fixed sample test. The Pocock (1982) boundaries offer a higher chance of early stopping even when the true treatment effect is near the initial estimate used to plan the trial. Furthermore, there are group sequential designs that fulfill certain criteria of optimality.

Summing up, adaptive designs are based on special combination rules for p-values such as Fisher's rule and do not offer such a large variety of possible plans as group sequential designs do. On the other hand, group sequential designs do not offer the flexibility to make data adaptive changes to the trial design during the course of the trial based on the results of interim analyses. In the present article, a method will be presented generalizing the data-adaptive approach to the large variety of group sequential designs by combination of both approaches. Particularly, at every interim analysis, the proposed method allows for a change of the sample size as well as of the alpha-spending function and the number and the time points of future interim analyses. As a typical example for the advantages of our proposed method, suppose that an extension of the trial turns out to be necessary in the last interim analysis. Just for the same ethical and economic aspects as when planning a new trial, one should include further interim analyses. This can be done with our method. See the example of a clinical trial in Parkinson's disease.

2. Representing a Group Sequential Test as a Combination Rule of Two Independent *p*-Values

In the past, adaptive designs were constructed by choosing a conventional combination rule, such as the commonly used Fisher's rule, for combination of the independent *p*-values derived from the stages of an experiment. The basic idea behind the method presented in this article is that every group sequential design implicitly defines a combination rule for *p*-values. At the time of the first interim analysis, the decision rule of a group sequential test can be represented as a combination rule for two *p*-values, one of the two *p*-values being derived from the data collected in the first stage of the trial and the other *p*-value being derived from the independent data collected in the further course of the trial, i.e., in stages 2 to m. Changes to the design can then be made following the first stage in the same manner as for adaptive designs based on conventional combination rules of *p*-values. By repetition, this principle can be applied to the further stages of the group sequential test as well because, at every time point of an interim analysis, the further stages of the trial can be understood as an independent new trial.

To define group sequential tests, let T_t denote the test statistic calculated on the basis of all data observed until the time point t. We assume that the stochastic process T_t has the information time parameter $t \in [0, 1]$ and follows a Brownian motion with drift parameter δ, which will be denoted by $T_t = B_{\delta t, t}$. Thus, T_t is a Gaussian process with expected value $\mathrm{E}(T_t) = \delta t$, variance $\mathrm{var}(T_t) = t$, and independent increments. By the central limit theorem and further theory for stochastic processes, many test statistics frequently used in clinical trials asymptotically follow a stochastic process that can be transformed into a Brownian motion. We refer to Müller and Schäfer (1999) for details and further references. See also the example below.

An m-stage group sequential test of the null hypothesis H_0: $\delta = 0$ is then given by time points $0 < t_1 < \cdots < t_m = 1$ and acceptance intervals $]a_1, b_1[, \ldots,]a_m, b_m[$ for the statistics $T_1, \ldots, T_m$, where the abbreviation $T_i = T_{t_i}$ is used. The null hypothesis is rejected in favor of H_1: $\delta > 0$ or H_1: $\delta < 0$ on stage k of the trial if $T_k \geq b_k$ or $T_k \leq a_k$, respectively. Rejection of H_0 causes a stop of the sequential procedure. One-sided tests of $\mathrm{H}_0^{\leq}$: $\delta \leq 0$ or $\mathrm{H}_0^{\geq}$: $\delta \geq 0$ are obtained by setting $a_i = -\infty$ or $b_i = +\infty$ for all $i = 1, \ldots, m$, respectively. One may calculate the cumulative rejection error probabilities $\alpha_{k,\mathrm{low}} = P_0(\mathrm{H}_0$ is rejected in favor of H_1: $\delta < 0$ up to stage $k)$ and $\alpha_{k,\mathrm{upp}} = P_0(\mathrm{H}_0$ is rejected in favor of H_1: $\delta > 0$ up to stage $k)$ by numerical integration with a convolution formula as used by Armitage, McPherson, and Rowe (1969) (see the Appendix). Then $\alpha = \alpha_{m,\mathrm{low}} + \alpha_{m,\mathrm{upp}}$ is the overall type I error risk of the group sequential test. Inversely, the time points and cumulative upper and lower rejection error probabilities determine corresponding boundaries $a_1, \ldots, a_m$ and $b_1, \ldots, b_m$.

We now want to represent this group sequential test as a combination rule of *p*-values at the time point of the first interim analysis. We are going to define the two *p*-values p and q based on the data of the first and the data of the further stages of the trial, respectively. For the sake of simplicity, we consider at first the case of a one-sided group sequential test of the null hypothesis $\mathrm{H}_0^{\geq}$: $\delta \geq 0$ versus $\mathrm{H}_1^{<}$: $\delta < 0$. And we use an unconventional definition of the *p*-value q that considerably simplifies the construction of the combination rule for p and q. Nevertheless, this definition of q leads to the same decision rule as with the usual definition of q following the method of Tsiatis, Rosner, and Metha (1984). The generalization to a two-sided test will be described at the end of this section. For a precise mathematical definition of the *p*-values p and q, we first define two random variables U and V by $U = T_1$ and $V = \min_{i=2,\ldots,m}(T_i - T_1 - a_i)$. U depends only on the data collected in the first stage of the trial, and V depends only on the data collected in the stages 2 to m. Since the increments $T_i - T_1$ are independent from T_1, the random variables U and V are stochastically independent. Let u, v, and $T_{i,\mathrm{obs}}$ denote realizations of the random variables U, V, and T_i, respectively. The decision rule defined by the group sequential test can be represented as a function of u and v. Since $T_{i,\mathrm{obs}} \leq a_i$ for some $i \geq 2$ is equivalent to $u + v \leq 0$, $\mathrm{H}_0^{\geq}$ will be rejected if and only if $u \leq a_1$ or $u + v \leq 0$.

Now let F_δ and G_δ denote the cumulative distribution functions of U and V, respectively, under the drift parameter δ. Let p and q denote the *p*-values corresponding to the realizations u and v of U and V, respectively, defined by $p = F_0(u)$ and $q = G_0(v)$. The *p*-values p and q are realizations of the random variables $F_0 \circ U$ and $G_0 \circ V$, respectively. Under the extreme value $\delta = 0$ of $\mathrm{H}_0^{\geq}$: $\delta \geq 0$, $F_0 \circ U$ and $G_0 \circ V$ follow a uniform distribution over the unit interval $[0, 1]$.

The decision rule defined by the group sequential test on the first stage can now be represented as a function of the *p*-values p and q. Under the assumptions described above, we have $p = F_0(u) = \Phi(u/{t_1}^{1/2})$ and $q = G_0(v)$, where Φ denotes the standard Gaussian distribution function. The calculation of G_0 is more complicated and needs numerical integration. This will be described in the Appendix. Applying the monotonically increasing functions F_0 and G_0, the inequalities $u \leq a_1$ and $v \leq -u$ derived above can be transformed into corresponding inequalities for p and q. The condition for the rejection of the null hypothesis $\mathrm{H}_0^{\geq}$: $\delta \geq 0$ becomes

$$p \leq \Phi\left(a_1/\sqrt{t_1}\right) \quad \text{or} \quad q \leq G_0\left\{\Phi^{-1}(1-p)\sqrt{t_1}\right\}. \qquad (*)$$

Here the symmetry $\Phi^{-1}(1-p) = -\Phi^{-1}(p)$ of the standard Gaussian distribution has been used.

The inequalities $(*)$ define a combination rule for the *p*-values p and q. Graphically, such a combination rule can be represented as an area in the (p, q) unit square. An example is shown in Figure 1. Condition $(*)$ defines a one-sided rejection region for rejection of the null hypothesis $\mathrm{H}_0^{\geq}$: $\delta \geq 0$, similar to the lower shadowed area in Figure 1. Actually, in Figure 1, the two-sided version of our method is plotted, which is presented

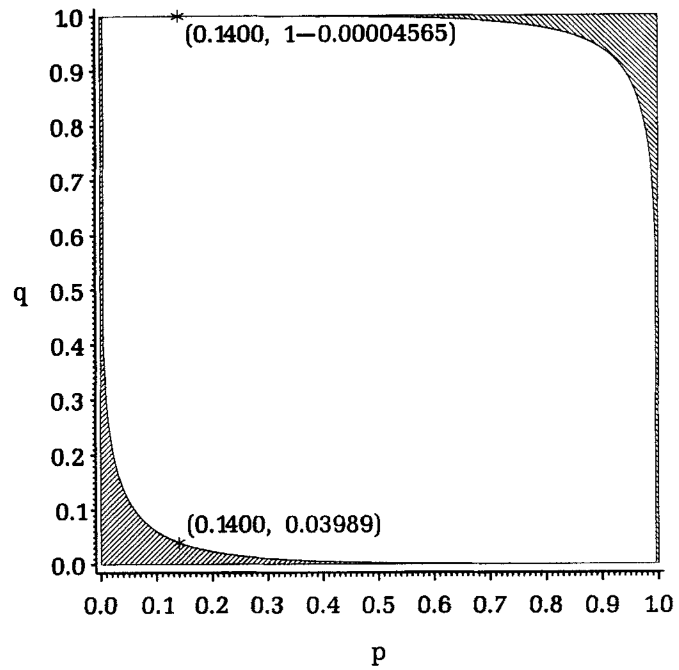

Figure 1. Representation of the initial group sequential design of the example as a combination rule of two p-values at the first interim analysis. p denotes the p-value calculated in the first interim analysis and q denotes the p-value derived from the further course of the trial, respectively, as defined in Section 2. The upper and the lower shadowed areas are the rejection regions for a two-sided test. After the interim analysis, the required significance level for the further course of the trial can be read from the boundary curve of the rejection regions. The marks $*$ show a supposed interim result with a p-value of 0.1400.

at the end of this section. (See also the example.) $H_0^{\geq}$ will be rejected if and only if the point (p, q), defined by the p-values of u (first stage of the trial) and v (further stages), lies in the lower shadowed area. The type I error level of the combination test defined by the combination rule of the two independent p-values is given by the area of this rejection region. This area is equal to the global type I error risk of the initially specified classical group sequential test. The boundary curve of a one-sided rejection region is defined by the function

$$\alpha(p) = \begin{cases} 1 & \text{for } p \leq \Phi\left(a_1/\sqrt{t_1}\right) \\ G_0\left\{\Phi^{-1}(1-p)\sqrt{t_1}\right\} & \text{for } p > \Phi\left(a_1/\sqrt{t_1}\right). \end{cases}$$

This α-function completely defines the testing procedure at the first stage of the trial in terms of p-values. If the p-value p derived from the data collected in the first stage of the trial is smaller than or equal to $\Phi(a_1/{t_1}^{1/2})$, then the pair of p-values (p, q) will lie in the lower shadowed rejection area, independently of the data collected in the further course of the trial, since $\alpha(p) = 1$. Hence, in this case, the trial will be stopped based on the first stage. Clearly, this exactly corresponds to the event that $T_1 \leq a_1$. If $p > \Phi(a_1/{t_1}^{1/2})$, the trial will be continued, and the summary p-value q obtained from the further stages of the trial must be smaller than or equal to $\alpha(p)$ in order to reject $H_0^{\geq}$. This just means that, in the rest of the trial, a level $\alpha(p)$-test of the null hypothesis $H_0^{\geq}$: $\delta \geq 0$ has to be performed in order to obtain an overall significance level equal to the global significance level of the initial group sequential test. Within this condition, the statistical test procedure applied during the further course of the trial may be redefined after the first interim analysis. The number and time points of further analyses as well as the sample size planned for the rest of the trial may be redesigned based on all the data observed in the first stage. The only condition is that the statistical test procedure for the further course of the trial holds level $\alpha(p)$ determined from the results of the first stage of the trial. Indeed, since the function $\alpha(p)$ itself is left unchanged and since this function completely determines the combination rule for the two independent p-values p and q, the type I error risk, which is equal to the area under the $\alpha(p)$-curve, also remains unchanged. The type I error risk does not depend on the statistical test by which the p-value q will be obtained. When applying the proposed method, the statistical test procedure for the further course of the trial has to be performed on data independent of the data up to the interim analysis and has to be designed without knowledge of the data derived from the further course of the trial.

If one does not want to change the design after the first interim analysis, then the rest of the group sequential design for the time points t_2 to t_m defines a group sequential level $\alpha(p)$-test for the rest of the trial. To understand this level $\alpha(p)$-test procedure as a test procedure for an independent new trial, the remaining part of the design can be transformed into a group sequential test for a Brownian motion starting at the time point of the interim analysis. The new time parameter is denoted by t' and the new drift parameter is given by $\delta' = \delta(1 - t_1)^{1/2}$. The time is reparameterized such that $t' = 0$ at $t = t_1$ and $t' = 1$ at $t = t_m = 1$, i.e., $t' = (t - t_1)/(1 - t_1)$. Indeed, $(B_{\delta t,t} - B_{\delta t_1,t_1})/(1 - t_1)^{1/2}$, where $t = t_1 + (1 - t_1)t'$ and $\delta = \delta'/(1 - t_1)^{1/2}$, defines a Brownian motion with time parameter t' and drift parameter δ'. The new testing times and the critical boundaries then are given by $t'_i = (t_{i+1} - t_1)/(1 - t_1)$ and $a'_i = (a_{i+1} - T_{1,\text{obs}})/(1 - t_1)^{1/2}$, respectively. This also shows that, if no design changes are necessary, then one can continue according to the initial plan and all the statistical properties having motivated the choice of this plan will be preserved.

The procedure described above may be applied again at any further stage of the trial. As explained before, the rest of the trial after the first interim analysis can be treated statistically as a new trial with an adjusted significance level $\alpha(p)$.

For a two-sided version of our method, one only has to replace G_0 by another function $\bar{G}_0$ in the definition of the α-function. $\bar{G}_0$ is defined by $\bar{G}_0(v) = P_0(R_j(v)$ for some $j \geq 2)$, where $R_j(v)$ denotes the event $\{T_j - T_1 - v \leq a_j$ and $T_i - T_1 - v \in]a_i, b_i[$ for all $2 \leq i < j\}$. This will define an α-function $\alpha_{\text{low}}(p)$ that is a boundary of a rejection region for rejecting H_0: $\delta = 0$ in favor of H_1: $\delta < 0$. Changing the roles of the lower boundaries a_i and the upper boundaries b_i in these definitions will result in an α-function $\alpha_{\text{upp}}(p)$ and a rejection region for rejecting H_0: $\delta = 0$ in favor of H_1: $\delta > 0$. This version is only slightly more efficient than simultaneously testing for the two one-sided alternatives with the presented procedure at nominal $\alpha/2$ levels. However, it automatically corrects for overlapping of the upper and lower parts of the rejection region, avoiding the possibility of contrary decisions. Note that both rejection regions are no longer symmetric where the boundaries of the sequential plan are asymmetric ($a_i \neq -b_i$ for some i).

Confidence intervals and point estimates following significance testing for the effect size parameter δ within the Brownian motion model may be obtained by combining the presented method with the method of Tsiatis et al. (1984). Indeed, there is a straightforward generalization of our method to simultaneously test null hypotheses of the form H_{0,δ_0}: $\delta = \delta_0$ also for δ_0 other than zero. This may be done by integrating the method of Tsiatis et al. (1984) for calculating p-values in group sequential tests. A confidence set will then consist of all parameter values δ_0 for which H_{0,δ_0} cannot be rejected at a significance level of one minus the specified confidence level. However, this is not the primary objective of the present paper.

3. Example

We are planning a clinical trial evaluating the effect of deep brain stimulation on motor function (part III of the Unified Parkinson's Disease Rating Scale, UPDRS III) and quality of life (Parkinson's Disease Questionnaire with 39 items, PDQ-39) in patients suffering from Parkinson's disease. In this trial, let n be the number of matched pairs to be recruited for randomization and let X_i be the difference in the change of quality of life of pair i (change, i.e., absolute difference of 6-month check-up value minus baseline value, in the summary index of PDQ-39 of the patient with deep brain stimulation minus the score of the patient without deep brain stimulation). Consider the partial sums $Z_k = \Sigma_{i=1}^{k} X_i$ of i.i.d. random variables X_i, $i = 1, \ldots, n$, every $X_i \sim \mathrm{N}(\mu, \sigma^2)$. Let $T_t = Z_{tn}/\sigma n^{1/2}$, where $t = k/n$ is the time parameter of the trial after the kth observation. Then T_t may be extended to a Brownian motion $B_{\delta t,t}$ with drift parameter $\delta = (\mu/\sigma)n^{1/2}$ (linkage formula).

The results of a trial of Martinez-Martin et al. (2000) on pallidotomy in patients with Parkinson's disease was used as a basis for the design of our trial, with the assumption that similar effects of deep brain stimulation could be expected. A gain in life quality of $\mu^* = -6$ points and a standard deviation of $\sigma = 17$ was derived from this study, leading to a value of $\mu^*/\sigma = -0.35$ for sizing the study. However, smaller values of $|\mu|$ were still considered clinically meaningful.

The type I error risk and the type II error risk for detection of this effect size were set to $\alpha = 0.05$ and $\beta = 0.2$, respectively. A classical group sequential design following Fleming et al. (1984) and spending a value of 0.01 of the type I error level for one interim analysis in the middle of the trial was chosen. In the Brownian motion model, $\mu^*/\sigma = -0.35$ corresponds to a value of the drift parameter of $\delta^* = -2.8293$ and the critical boundaries for T_t are ± 1.8214 at $t = t_1 = 0.5$ and ± 2.0027 at $t = t_2 = 1$. Using the linkage formula above, a maximum number of $n = n_{\max} = \delta^{*2}/(\mu^*/\sigma)^2 = 2.8293^2/0.35^2 = 66$ matched pairs have to be recruited for randomization, only one more than with the fixed sample test. Under $\mu = \mu^*$, the average number of matched pairs to be assessed is 56, nine less than with the fixed sample test. The first interim analysis is planned after assessment of 33 matched pairs.

Since a real consensus for specification of the minimal clinically relevant difference was missing and since there are uncertainties in the choice of the detectable difference due to an uncertain estimation of σ, the method for redesigning the trial as proposed in the present article was included in the trial protocol. By the way, the implementation of the method in the study protocol also will offer the possibility to replace the main outcome variable. In the case that the summary index of PDQ-39 was modified during the course of the trial or that a more sensitive quality of life index in Parkinson's disease is available at the time of an interim analysis, we can decide about a change of the primary outcome variable.

The p-value representation of the group sequential design at the first interim analysis is shown in Figure 1. The lower shadowed area is the rejection region for H_0: $\delta = 0$ in favor of H_1: $\delta < 0$. The boundary curve of the lower shadowed area is given by the function $\alpha(p)$. The upper shadowed area defines the corresponding rejection region for H_0: $\delta = 0$ in favor of H_1: $\delta > 0$.

As a numerical example, suppose that in the Parkinson's disease trial σ was estimated by $s = 26.91$ and μ by $\bar{x} = -167/33 = -5.061$ based on the assessment of the first 33 matched pairs at interim. Then $T_{1,\mathrm{obs}} = -0.7639$ is the observed value of the test statistic, which corresponds to a nominal one-sided p-value of 0.1400. The point $(p, 1.0)$ does not lie in the rejection region; hence, H_0 cannot be rejected. The trial must be continued. From the lower boundary curve in Figure 1, the nominal α-level $\alpha'_{\mathrm{low}} = 0.03989$ is obtained for planning the further stages of the trial for testing H_0: $\delta = 0$ against H_1: $\delta < 0$. The corresponding significance level $\alpha'_{\mathrm{upp}} = 0.00004565$ for testing against H_1: $\delta > 0$ is obtained from the upper boundary curve. These significance levels are determined by numerical integration as described in the Appendix. Replacing the density functions $f_{0,i}$ by the respective density functions $f_{\delta,i}$ under the special value $\delta = -\delta^*$ or $\delta = +\delta^*$ of the drift parameter, one may also calculate the conditional power to detect a treatment difference of $-\delta^*$ or of $+\delta^*$ in the rest of the trial. The results are 0.5982 and 0.02794, respectively. $\delta^* = -2.8293$ transforms into $\delta'^* = -2.0006$.

Suppose there is consensus in the study group that a value of $\mu'^*/\sigma = -0.2$, which is the result of rounding the observed value of $\bar{x}/s = -0.1881$, is still worthwhile to be detected in the rest of the trial. However, without an extension of the sample size, the conditional power to detect this treatment difference is only 0.2713. It is decided to extend the trial to reach a conditional power of 0.8 under $\mu'^*/\sigma = -0.2$. For a fixed sample design with the one-sided α-level $\alpha'_{\mathrm{low}} = 0.03989$, recalculation of the sample size would give $(\delta'^*_{\mathrm{fixed}})^2/(\mu'^*/\sigma)^2 = 169$ matched pairs by use of $\delta'^*_{\mathrm{fixed}} = -\{\Phi^{-1}(1 - 0.03989) + \Phi^{-1}(0.8)\} = -(1.7520 + 0.8416) = -2.5936$. Since this is a fairly large extension, it is decided to include further interim analyses. With existing methods, however, a design extension can only be made without inclusion of further interim analyses. In contrast, with our method, we chose a three-stage group sequential design for reduction of the average sample size and for early stopping if $\mu/\sigma = \mu^*/\sigma = -0.35$ instead of $\mu/\sigma = \mu'^*/\sigma = -0.2$.

This group sequential test is chosen in the spirit of Fleming et al. (1984), spending lower α-values 0.01, 0.02, and 0.03989 and upper α-values 0.00001, 0.00002, and 0.00004565 at the information times $t'_1 = 0.3333$, $t'_2 = 0.6667$, and $t'_3 = 1$, respectively. This transforms into lower and upper critical boundaries $a'_1 = -1.3431$, $a'_2 = -1.8121$, $a'_3 = -1.8914$, and $b'_1 = 2.4624$; $b'_2 = 3.4704$, $b'_3 = 4.0236$, respectively, for the analyses of a Brownian motion in the time parameter t' and drift parameter δ'. With this plan, the desired power of 0.8 is reached at $\delta'^* = -2.6819$. Only a small increase of the maximum sample size to 180 has to be paid with the advantage that the average sample size is only 136 at $\mu/\sigma = \mu'^*/\sigma = -0.2$. Under $\mu/\sigma = \mu'^*/\sigma = -0.35$, there is a chance of 0.6493 and 0.9508 to stop the trial with claimed advantage of the deep brain stimulation at the first and at the first or second interim analysis, respectively. Thus, the average sample size is only 84 under this effect size parameter that was used for planning the trial at the beginning.

4. Discussion

A method was presented for a full integration of the concept of adaptive interim analyses (Bauer and Köhne, 1994) into group sequential testing. This new methodology is analogous to considering data from before and after an interim analysis point as two separate studies. Our method is the first method allowing for a change of the alpha-spending function and the number and the time points of future interim analyses. This introduces the flexibility of adaptive trial designing into the wide world of group sequential plans. Any group sequential plan can be given the full flexibility of an adaptive design. Data adaptive design changes can be made after every interim analysis. From the other point of view, this method enlarges the class of adaptive designs beyond the commonly used combination rules such as Fisher's combination rule or its modifications. Adaptive designs can now be chosen from the large class of group sequential designs and can better be fitted to special situations. When no changes to the design are necessary during the course of the trial, the trial can be continued on the basis of the initially planned group sequential test, with the consequence that all the statistical properties for which the special group sequential design was selected remain valid. Especially in classical group sequential testing, it is possible to choose an optimal group sequential design with respect to a specified optimization criterion such as minimal average sample size and to make this design fully adaptive in the sense of Bauer and Köhne (1994). The method requires numerical integration, which should not be a problem with modern computers.

Of course, a decision to change a design following an interim analysis always needs careful consideration. In our view, the most important application of our method is the adjustment of the sample size and the number and time points of interim analyses. Although it is also possible to change the test statistic or even the primary outcome variable, such changes should be restricted to exceptional situations. In these cases, one has to pay attention to the problem of the interpretation of the final results with respect to the clinical claims. Careful attention must also be given to potential sources of operational bias. Fortunately, the necessity to change a primary outcome measure is rare, e.g., if recommendations from (regulatory) authorities changed during a long-term study. Changing the test statistic is of minor interest when comparing two treatments but may become important in trials with more than two treatment groups, e.g. in dose-finding studies.

The proposed method requires careful application and prespecifications in the protocol. The function $\alpha(p)$ defined by the chosen group sequential plan at the first interim analysis should be specified unambiguously in the protocol in addition to the usual specifications for the group sequential design. After every design change, an amendment to the study protocol should immediately be made, making the corresponding specifications for the rest of the trial, including the new group sequential plan as well as the new function $\alpha(p)$. The amendment should also explain the reasons for changing the design.

Exact p-values instead of asymptotic ones may be used for the decision rule defined by the α-function. The type I error will be preserved under the condition that the p-values are uniformly distributed over the unit interval. Discrete distributions will work with slight modifications. However, other properties such as characteristic curves of power and average sample size will hold only approximately where the statistical model is asymptotically a Brownian motion model.

Fisher (1998) presented a different method for sample size adjustment after an interim analysis. His method for self-designing clinical trials is based on a variance-spending approach. The type I error level is preserved by use of a weighted cumulative statistic. The method, however, does not include the possibility of early stopping with rejection of the null hypothesis at an interim analysis but only early stopping for futility. Based on the approach by Fisher (1998), Shen and Fisher (1999) proposed an automatic algorithm for the sample-size adjustment. In their algorithm, the sample size is a function of the point estimate of the effect size at an interim analysis. Our method can easily be extended in this direction, and both methods may also use confidence limits of the effect size estimate instead of a point estimate. However, we consider it as an advantage of our method that there is no automatism for design changes based on an interim result. All the data collected up to the interim analysis as well as external information can be taken into account when redesigning the trial. Particularly, all information and assumptions having led to the choice of the initial design may be reassessed.

Résumé

Nous présentons une méthode générale qui intègre le concept d'analyses intermédiaires flexibles dans les tests séquentiels groupés classiques. Le chercheur peut ainsi représenter chaque plan séquentiel groupé comme un plan expérimental flexible, et modifier ce plan, en cours d'essai, après chaque analyse intermédiaire, comme pour un plan flexible. Le concept de plan d'essai flexible est généralisé ainsi à de nombreux plans séquentiels possibles.

References

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

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. (1992). The choice of sequential boundaries based on the concept of power spending. *Biometrie und Informatik in Medizin und Biologie* **23,** 3–15.

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

Brittain, E. H. and Bailey, K. R. (1993). Optimization of multistage testing times and critical values in clinical trials. *Biometrics* **49,** 763–772.

Cui, L., Huang, H. M. J., and Wang, S.-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55,** 853–857.

DeMets, D. L. and Ware, J. H. (1980). Group sequential methods in 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.

Fisher, L. D. (1998). Self-designing clinical trials. *Statistics in Medicine* **17,** 1551–1562.

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

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

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculation in group sequential trials. *Biometrics* **55,** 1286–1290.

Martinez-Martin, P., Valldeoriola, F., Molinuevo, J. L., Nobbe, F. A., Rumia, J., and Tolosa, E. (2000). Pallidotomy and quality of life in patients with Parkinson's disease: An early study. *Movement Disorders* **15,** 65–70.

Müller, H.-H. and Schäfer, H. (1999). Optimization of testing times and critical values in sequential equivalence testing. *Statistics in Medicine* **18,** 1769–1788.

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.

Shen, Y. and Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55,** 190–197.

Tsiatis, A. A., Rosner, G. L., and Metha, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40,** 797–803.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43,** 193–200.

Wassmer, G. (1997). A technical note on the power determination for Fisher's combination test. *Biometrical Journal* **39,** 831–838.

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

Received August 1999. *Revised January* 2001.
Accepted January 2001.

Appendix

Numerical Calculations

The numerical evaluation of the function $\alpha(p)$ is essential for the application of the procedure. By definition of $\alpha(p)$, we have to calculate $G_0(v)$ at $v = \Phi^{-1}(1-p){t_1}^{1/2}$. It is easier to calculate the complementary value $1 - G_0(v) = P_0(V > v) = P_0(T_i - T_1 > a_i + v \text{ for all } i = 2, \ldots, m)$.

With the notations and assumptions introduced in Section 2, let $f_{\delta,i}(x)$, $i = 2, \ldots, m$, denote the probability density functions of the increments $T_i - T_{i-1}$. These independent random variables follow Gaussian distributions with expected values $\delta(t_i - t_{i-1})$ and variances $t_i - t_{i-1}$. Hence,

$$f_{\delta,i}(x) = \frac{1}{\sqrt{2\pi}\sqrt{t_i - t_{t-1}}} \times \exp\left[-\frac{1}{2}\,\frac{\{x - \delta\,(t_i - t_{i-1})\}^2}{t_i - t_{i-1}}\right] \quad \text{for } i = 2, \ldots, m.$$

The value of $1 - G_0(v)$ can now be obtained by numerical integration based on an iterative application of the following convolution formula. We define functions $g_{\delta,i}(x)$ for $i = 2, \ldots, m$ by $g_{\delta,2}(x) = f_{\delta,2}(x)$ for $i = 2$ and, recursively, $(i - 1 \rightarrow i)$, $g_{\delta,i}(x) = \int_{a_{i-1}+v}^{\infty} f_{\delta,i}(x-y) g_{\delta,i-1}(y)dy$ for $i > 2$. Then $1 - G_0(v) = \int_{a_m+v}^{\infty} g_{0,m}(x)dx$, applying the formulas for $\delta = 0$. Hence, the rejection error level $\alpha(p)$ for the further course of the trial is determined. Conditional power values may be calculated in the same way using the formulas for an arbitrary δ. Time for calculating a conditional power characteristic curve or other characteristics depending on the parameter δ may be reduced by use of the identity

$$g_{\delta,i}(x) = g_{0,i}(x) \times \exp\left[\delta\left\{x - \frac{1}{2}\delta\,(t_i - t_1)\right\}\right] \quad \text{for } i = 2, \ldots, m.$$