

Adaptive Group Sequential Confidence Intervals for the Ratio of Normal Means

Joachim Hartung¹ and Guido Knapp

Department of Statistics, Dortmund University of Technology, Dortmund, Germany

Abstract: In controlled group sequential trials, we consider the ratio of normal means as the effect measure of interest and derive group sequential nested confidence intervals on this parameter. In an interim analysis, the sample sizes of the following stages can be determined in a completely adaptive way using the unblinded data from all previously performed stages. Despite the data-dependent adaptation, the nested confidence intervals always keep the predefined confidence coefficient.

Using nested confidence intervals, we make test decisions either in noninferiority or in superiority trials. However, in an interim analysis, we can change the planning from showing noninferiority to showing superiority or vice versa without affecting the predefined confidence coefficient of the nested intervals.

A real data example is worked out in detail and the change in the planning from showing superiority to showing noninferiority is shown during the ongoing trial.

Keywords: Ratio of means; Adaptive sample size planning; Switching between noninferiority and superiority; Group sequential clinical trials

1 Introduction

In clinical trials, the difference or the ratio of normal means are the common effect measures for showing at least noninferiority of a new treatment compared to a standard (active) treatment. Provided the standard treatment is well known and stable in different populations, the suitable effect measure is the difference of means. Otherwise, the scale invariant ratio of means is the preferred effect measure.

In analyzing the trials, the confidence interval approach is of particular attractiveness. Group sequential confidence intervals for the difference of means are known as repeated confidence intervals, see Jennison and Turnbull (2000) and references cited therein, Hartung and Knapp (2006).

¹Address correspondence to Joachim Hartung, Department of Statistics, Dortmund University of Technology, 44221 Dortmund, Germany; E-mail: hartung@statistik.tu-dortmund.de

A well-known confidence interval for the ratio of normal means was derived by Fieller (1940). Fieller parameterized a F -statistic, where the ratio of the means, say λ , is the parameter of interest. The boundaries of the confidence interval are found by equating the F -statistic to quantiles of an appropriate F -distribution and solving for the unknown parameter. The boundaries are given by solving a quadratic equation, see also Finney (1964). In spite of its practical importance, the ratio of means does not seem to be considered in group sequential trials until now, neither for testing noninferiority nor for deriving confidence intervals.

In this paper, we will extend the proceeding of Fieller (1940) to group sequential trials that are designed for a fixed number, say K , of independent stages, where the ratio of normal means is the parameter of interest. We consider pivotal t -statistics in each stage involving the unknown parameter λ and combine the pivotal statistics using the inverse normal method from meta-analysis, see for instance Hartung, Knapp, and Sinha (2008). As with Fieller's confidence interval, the multi-stage confidence intervals are defined implicitly and, for obtaining the boundaries, nonlinear equations have to be solved. However, the solutions are always unique.

In each stage k , $1 \leq k \leq K$, a confidence interval on λ will be computed using the data of all previous stages. The consecutive intersection of these individual confidence intervals leads to a sequence of intervals that are nested. This property is of particular interest in the confidence interval approach to the analysis of noninferiority trials, see Bauer and Kieser (1996) and, for instance, the clinical trial guideline EMEA (2000). Practically this means that the position of the confidence interval determines the kind of result of the study, independently of the question, whether the study has been originally planned as a noninferiority or as a superiority trial. The consequence of the proposed confidence interval intersection-approach is that, if we gain significance for noninferiority at an early stage of the trial, we will not take a risk to lose this significance in case we decide to continue the trial for an attempt of showing superiority.

In group sequential trials, interim analyses are based on the unblinded data. Since stochastically independent and uniformly distributed p -values will be combined for constructing the confidence intervals, the information from the interim analyses of the previous stages may be used for an adaptive sample size calculation of the following stage, see Brannath, Posch, and Bauer (2002), Hartung (2006). We will provide concrete rules for updating sample sizes during the course of the group sequential trial.

The outline of the present paper is as follows: In Section 2.1, one-sided group sequential

confidence intervals for the ratio of normal means are derived. Moreover, we consider the change from showing noninferiority to showing superiority during the course of the group sequential trial. In Section 2.2, two-sided confidence intervals and a test on the homogeneity of the ratios of the mean treatment effects underlying the K independent stages are presented. A group sequential confidence interval on the variance parameter is given in Section 3. General adaptive sample size planning is considered in Section 4. In Section 5, a real data example is presented that uses an adaptive design of O'Brien and Fleming (1979) type. In the example, the switching from showing superiority to showing noninferiority is shown during the ongoing trial. Moreover, after the first interim analysis, the design was changed from a three-stage into a two-stage design and we will show how to make full use of the type I error rate despite the shortening of the originally planned trial, that is, we will not lose any power. Some final remarks are given in Section 6, where especially point estimation of the ratio λ is briefly addressed.

2 Group Sequential Confidence Intervals for the Ratio of Normal Means

Let X_E and X_C be stochastically independent normally distributed random variables with mean $\mu_E \geq 0$ in an experimental group E, mean $\mu_C > 0$ in an active control group C, and common variance $\sigma^2 > 0$, succinctly, $X_E \sim \mathcal{N}(\mu_E, \sigma^2)$ and $X_C \sim \mathcal{N}(\mu_C, \sigma^2)$. The parameter of interest is the ratio of the means, say λ , that is,

$$\lambda = \frac{\mu_E}{\mu_C}, \quad 0 \leq \lambda < \infty, \quad \mu_E \geq 0, \quad \mu_C > 0.$$

Let Δ_0 , with $0 \leq \Delta_0 < 1$, be a noninferiority margin, we are interested in hypothesis testing of

$$H_{0,\Delta} : \lambda = 1 - \Delta \quad \text{versus} \quad H_{1,\Delta} : \lambda > 1 - \Delta, \quad 0 \leq \Delta \leq \Delta_0, \quad (1)$$

at a predefined level α , $0 < \alpha < 1/2$. The alternative hypothesis $H_{1,\Delta}$ stands for (Δ -)noninferiority, $0 < \Delta \leq \Delta_0$, and means superiority of the experimental group E with regard to the control group C for $\Delta = 0$.

We consider a comparative study which is carried out consecutively in a number of independent stages, say K . In the i -th stage, $i = 1, \dots, K$, let be $\bar{X}_{E_i} \geq 0$ and $\bar{X}_{C_i} > 0$ the sample means of $n_{E_i} \geq 2$ and $n_{C_i} \geq 2$ responses in the respective treatment groups. The variance parameter σ^2 is estimated in the i -th stage by the pooled sample variance

S_i^2 , which is stochastically independent of the means and follows a scaled χ^2 -distribution with $n_{E_i} + n_{C_i} - 2$ degrees of freedom, that is

$$(n_{E_i} + n_{C_i} - 2) \frac{S_i^2}{\sigma^2} \sim \chi_{n_{E_i} + n_{C_i} - 2}^2. \quad (2)$$

2.1 Nested One-Sided Confidence Intervals

For the i -th stage, $i = 1, \dots, K$, let us introduce the pivotal statistic

$$\bar{X}_i(\lambda) = \bar{X}_{E_i} - \lambda \bar{X}_{C_i} \sim \mathcal{N}\left(0, \left(\frac{1}{n_{E_i}} + \frac{\lambda^2}{n_{C_i}}\right) \sigma^2\right), \quad \bar{X}_{E_i} \geq 0, \quad \bar{X}_{C_i} > 0, \quad (3)$$

and define the parameterized pivotal t -statistic

$$T_i(\lambda) = \frac{\bar{X}_{E_i} - \lambda \bar{X}_{C_i}}{S_i \sqrt{1/n_{E_i} + \lambda^2/n_{C_i}}} \sim t_{n_i - 2}, \quad n_i = n_{E_i} + n_{C_i}. \quad (4)$$

That is, for the *true* parameter λ , the pivotal statistic $T_i(\lambda)$ from (4) follows a central t -distribution with $n_{E_i} + n_{C_i} - 2$ degrees of freedom.

Let F_{t_ν} denote the cumulative distribution function of a t -variable with ν degrees of freedom, then it holds, for the $1 - p$ -value,

$$F_{t_{n_i - 2}}(T_i(\lambda)) \sim U(0, 1), \quad n_i = n_{E_i} + n_{C_i}, \quad (5)$$

where $U(0, 1)$ stands for the uniform distribution in the unit interval. Consequently, we obtain

$$\Phi^{-1}[F_{t_{n_i - 2}}(T_i(\lambda))] \sim \mathcal{N}(0, 1), \quad i = 1, \dots, K, \quad (6)$$

with Φ^{-1} the inverse of the standard normal cumulative distribution function Φ .

The stages of the trial are assumed to be independent. Consequently, up to the j -th stage, we define the combining pivotal statistic

$$Z_j(\lambda) = \sum_{i=1}^j \Phi^{-1}[F_{t_{n_i - 2}}(T_i(\lambda))] \sim \sqrt{j} \mathcal{N}(0, 1), \quad j = 1, \dots, K. \quad (7)$$

Let Y_1, \dots, Y_K , in general, be mutually independent $\mathcal{N}(0, 1)$ -distributed random variables. Then, for predefined level α , $0 < \alpha < 1/2$, positive critical values cv_1, \dots, cv_K may be defined which satisfy the following probability statement,

$$P\left(\sum_{i=1}^j Y_i \leq cv_j \text{ for all } j = 1, \dots, K\right) = 1 - \alpha. \quad (8)$$

For fixed number of stages K and overall significance level α , we get an O'Brien and Fleming (1979) design with constant critical values in (8), say $cv_j = cons_{OBF}(K, \alpha)$, and a Pocock (1977) design with monotone increasing critical values given as $cv_j = \sqrt{j} cons_{PO}(K, \alpha)$, $j = 1, \dots, K$, see Hartung (2006).

Note that upper and lower critical values at level 2α for the corresponding symmetric two-sided tests are usually tabulated in the literature, which, for $K \geq 2$, are slightly closer to zero than the one-sided critical values at level α . But at least for $\alpha \leq 0.05$, these upper and lower two-sided critical values may be used as critical values in one-sided test problems for most practical applications, see Jennison and Turnbull (2000, p. 192).

Using critical values cv_j defined by (8), we get the following probability statements for the combining pivotal statistic from (7),

$$P_\lambda \left(Z_j(\lambda) \leq cv_j \text{ for } j = 1, \dots, k \leq K \right) \begin{cases} \geq 1 - \alpha & \text{for } k < K, \\ = 1 - \alpha & \text{for } k = K. \end{cases} \quad (9)$$

Definition 1. *The group sequential lower confidence sets on the ratio λ are defined as*

$$CI_{k,L}(\lambda) = \{ \tilde{\lambda} \geq 0 \mid Z_j(\tilde{\lambda}) \leq cv_j \text{ for } j = 1, \dots, k \}, \quad k = 1, \dots, K. \quad (10)$$

By (9), we can directly state the following lemma:

Lemma 1. *The lower confidence sets from Definition 1 are nested, that is,*

$$CI_{k+1,L}(\lambda) \subset CI_{k,L}(\lambda), \quad k = 1, \dots, K - 1, \quad (11)$$

and the confidence coefficient of $CI_{k,L}(\lambda)$ is at least $1 - \alpha$, and exactly $1 - \alpha$ for $k = K$.

In order to show that these confidence sets are genuine intervals, we need a further lemma.

Lemma 2. *The statistics $Z_j(\lambda)$, $j = 1, \dots, K$, from (7) are monotone decreasing in λ , for $0 \leq \lambda < \infty$.*

Proof. At first, we show that the pivotal t-statistics $T_i(\lambda)$, $i = 1, \dots, K$, are (strictly) monotone decreasing in λ for $\lambda \geq 0$.

Suppressing the subscript i and putting $Q = (1/n_E + \lambda^2/n_C) S^2)^{1/2}$, we get the

derivative

$$\begin{aligned}
\frac{d}{d\lambda} T(\lambda) &= \frac{-\bar{X}_C Q - (\bar{X}_E - \lambda \bar{X}_C) Q^{-1} S^2 \lambda / n_C}{Q^2} \\
&= \frac{-\bar{X}_C Q^2 - (\bar{X}_E - \lambda \bar{X}_C) S^2 \lambda / n_C}{Q^3} \\
&= \frac{-(\bar{X}_C / n_E + \lambda \bar{X}_E / n_C) S^2}{Q^3} < 0 \quad \text{for } \lambda \geq 0,
\end{aligned} \tag{12}$$

so that $T(\lambda)$ is monotone decreasing for nonnegative λ . Since the functions $F_{t_\nu}(\cdot)$ and $\Phi^{-1}(\cdot)$ used in building $Z_j(\lambda)$, see (7), are both (strictly) monotone increasing functions in their arguments, the function $\Phi^{-1}(F_{t_\nu}(T(\lambda)))$ is monotone decreasing in $\lambda \geq 0$. Hence the combined statistic $Z_j(\lambda)$ from (7) is monotone decreasing in λ for $\lambda \geq 0$. \square

Moreover, the combining pivotal statistic $Z_j(\lambda)$ is a bounded function with respect to $\lambda \geq 0$, that is, see (7), for $j = 1, \dots, K$,

$$\begin{aligned}
\sum_{i=1}^j \Phi^{-1} \left[F_{t_{n_i}-2} \left(\frac{\bar{X}_{E_i}}{\sqrt{S_i^2 / n_{E_i}}} \right) \right] &= Z_j(0) \geq Z_j(\lambda) > \inf_{\lambda > 0} Z_j(\lambda) \\
&= \sum_{i=1}^j \Phi^{-1} \left[F_{t_{n_i}-2} \left(\frac{-\bar{X}_{C_i}}{\sqrt{S_i^2 / n_{C_i}}} \right) \right] =: Z_j(\infty).
\end{aligned} \tag{13}$$

Let us define, in the j -th stage, the following individual lower confidence set for λ ,

$$I_{j,L}(\lambda) = \{\tilde{\lambda} \geq 0 \mid Z_j(\tilde{\lambda}) \leq cv_j\}, \quad j = 1, \dots, K. \tag{14}$$

Then, by Lemma 2, the confidence set (14) is a connected lower bounded one-sided confidence interval on λ . This interval can also be described as

$$I_{j,L}(\lambda) = [\lambda_L(j), \infty), \tag{15}$$

where $\lambda_L(j)$ solves: $Z_j(\lambda_L(j)) = cv_j$ if $Z_j(0) > cv_j$, otherwise $\lambda_L(j) = 0$, $j = 1, \dots, K$. Note that the solution in (15) is unique and can be easily found iteratively using, for instance, the bisection method.

Since $CI_{k,L}$ from (10) is the intersection of the individual confidence intervals from (14) up to stage k , that is,

$$CI_{k,L}(\lambda) = I_{1,L}(\lambda) \cap \dots \cap I_{k,L}(\lambda), \quad k = 1, \dots, K, \tag{16}$$

we can summarize the above findings as follows.

Theorem 1 *The lower confidence interval on λ with a confidence coefficient of at least $1 - \alpha$, $0 < \alpha < 1/2$, is given in the k -th stage, $k = 1, \dots, K$, by*

$$CI_{k,L}(\lambda) = [\lambda_{Lk}, \infty) \quad (17)$$

where $\lambda_{Lk} = \max\{\lambda_L(1), \dots, \lambda_L(k)\}$ with $\lambda_L(j)$ from (15), $j = 1, \dots, k$. For $k = K$, the confidence coefficient is exactly $1 - \alpha$.

Let us now apply these group sequential confidence intervals to the test problem (1) at level of at most α . At stage k , $k = 1, \dots, K$, using λ_{Lk} from (17), we proceed as follows:

if $1 - \Delta < \lambda_{Lk}$, then we decide for $H_{1,\Delta}$, $\Delta \in [0, \Delta_0]$,

if $1 - \Delta_0 \geq \lambda_{Lk}$, then we stay with H_{0,Δ_0} .

If we are satisfied with showing noninferiority, then we will stop the trial after that stage k^* , when the value $1 - \Delta_0$ lies the first time outside the corresponding confidence interval. Since the confidence intervals $CI_{k,L}$ are nested and provided $k^* < K$, we may decide to continue the trial without any risk to lose the noninferiority once shown. In case unexpected favorable estimates of the involved parameters have been observed up to stage k^* , this may lead to considerations to switch from showing noninferiority to showing superiority. The trial is then continued by planning with $\Delta = 0$.

Conversely, originally planned as a superiority trial, a first interim analysis may reveal that an unexpected high number of subjects would be required for showing superiority. So, in case of an active control, one may decide to switch from showing superiority to showing noninferiority, and to reduce the sample size of the rest of the trial by choosing some $\Delta > 0$ in the further planning. Note that also in this situation, a noninferiority bound Δ_0 should have been defined at the beginning of the study, see also the discussion in the clinical trial guideline EMEA (2000).

2.2 Nested Two-Sided Confidence Intervals and Homogeneity of the Ratios of the Treatment Effects

Let us define in analogy to (10) the upper confidence sets for λ as

$$CI_{k,U}(\lambda) = \{\tilde{\lambda} \geq 0 \mid -cv_j \leq Z_j(\tilde{\lambda}) \text{ for } j = 1, \dots, k\}, \quad k = 1, \dots, K. \quad (18)$$

Again, the confidence sets are nested, that is, $CI_{k+1,U}(\lambda) \subset CI_{k,U}(\lambda)$, $k = 1, \dots, K - 1$, and each confidence set has a confidence coefficient of at least $1 - \alpha$ and exactly $1 - \alpha$ for $k = K$.

Definition 2. *The group sequential two-sided confidence set for λ in stage k , say $CI_k(\lambda)$, is defined as the intersection of the lower and the upper confidence sets from (10) and (18), respectively, that is,*

$$CI_k(\lambda) = CI_{k,L}(\lambda) \cap CI_{k,U}(\lambda), \quad k = 1, \dots, K. \quad (19)$$

Let $I_{j,U}(\lambda) = \{\tilde{\lambda} \geq 0 \mid -cv_j \leq Z_j(\tilde{\lambda})\}$ denote the individual upper confidence set in the j -th stage. In analogy to (15), we get the following interval representation,

$$I_{j,U}(\lambda) = [0, \lambda_U(j)], \quad (20)$$

where $\lambda_U(j)$ uniquely solves: $Z_j(\lambda_U(j)) = -cv_j$ if $Z_j(\infty) < -cv_j$, otherwise $\lambda_U(j) = \infty$. So the upper confidence set $CI_{k,U}$ from (18) has the interval representation,

$$CI_{k,U}(\lambda) = [0, \lambda_{Uk}], \quad k = 1, \dots, K, \quad (21)$$

where $\lambda_{Uk} = \min\{\lambda_U(1), \dots, \lambda_U(k)\}$ with $\lambda_U(j)$ from (20), $j = 1, \dots, k$.

Summarizing the above considerations, we can formulate the following results.

Theorem 2. *The two-sided confidence set $CI_k(\lambda)$ has the interval representation*

$$CI_k(\lambda) = [\lambda_{Lk}, \lambda_{Uk}], \quad k = 1, \dots, K, \quad (22)$$

where λ_{Lk} is from (17) and λ_{Uk} is from (21). The confidence intervals are nested,

$$CI_{k+1}(\lambda) \subset CI_k(\lambda), \quad k = 1, \dots, K - 1,$$

and each confidence interval has a confidence coefficient of at least $1 - 2\alpha$, $0 < \alpha < 1/2$.

Remark 1. Denote $I_k = [\lambda_L(k), \lambda_U(k)]$, see (15) and (20), the individual two-sided confidence interval in the k -th stage, then $CI_1 = I_1$ and $CI_k = CI_{k-1} \cap I_k$, $k = 2, \dots, K$.

Depending on the choice of α , the two-sided confidence interval CI_k from (22) may be empty, that is, it may occur that $\lambda_{Uk} < \lambda_{Lk}$. For interpreting such an event, let us consider the extended model that each stage has an individual parameter, say $\lambda_i = \mu_{E_i}/\mu_{C_i}$, $i = 1, \dots, K$. Let $b' = (b_1, \dots, b_k)$ denote the transposed of a vector b in the nonnegative orthant \mathbb{R}_+^k of \mathbb{R}^k , then, by (6) and (8), the k -dimensional confidence region,

$k = 1, \dots, K,$

$$\begin{aligned}
CR_k &= \left\{ b \in \mathbb{R}_+^k \mid -cv_j \leq Z_j(b_1, \dots, b_j) \right. \\
&:= \left. \sum_{i=1}^j \Phi^{-1} \left[F_{t_{n_{E_i} + n_{C_i} - 2}} \left(\frac{\bar{X}_{E_i} - b_i \bar{X}_{C_i}}{\sqrt{(1/n_{E_i} + b_i^2/n_{C_i}) S_i^2}} \right) \right] \leq cv_j \text{ for } j = 1, \dots, k \right\} \quad (23)
\end{aligned}$$

covers the parameter vector $(\lambda_1, \dots, \lambda_k)'$ with the probability of at least $1 - 2\alpha$, $0 < \alpha < 1/2$. Note that CR_k is not empty for all $\alpha \in (0, 1/2)$, as, for instance, the realized vector $(\bar{x}_{E_1}/\bar{x}_{C_1}, \dots, \bar{x}_{E_k}/\bar{x}_{C_k})'$ always lies within CR_k .

When we assume that the parameters λ_i are really identical, say $\lambda_i = \lambda$, $i = 1, \dots, k$, then the k -dimensional parameter vector $(\lambda, \dots, \lambda)'_k$ is covered by CR_k , or, in other words, $(\lambda, \dots, \lambda)'_k \in CR_k$ with probability of at least $1 - 2\alpha$. But this is equivalent to $\lambda \in CI_k$ with probability of at least $1 - 2\alpha$. Thus, if CI_k is empty for a common confidence level $1 - 2\alpha$, this will speak against the assumption of an identical ratio of mean treatment effects over the first k stages. This can formally be stated as a test on homogeneity of the ratios.

In testing

$$H_{0,hom}(k) : \lambda_1 = \dots = \lambda_k \quad \text{versus} \quad H_{1,hom}(k) : \lambda_{i_1} \neq \lambda_{i_2}, \quad (24)$$

for some $i_1, i_2 \in \{1, \dots, k\}$, $k = 2, \dots, K$, the homogeneity hypothesis $H_{0,hom}(k)$ will be rejected at level of at most 2α , if the confidence interval CI_k from (22) is empty. If $H_{0,hom}(k^*)$ is rejected, then also $H_{0,hom}(k)$ will be rejected for $k^* \leq k \leq K$. An alternative to this homogeneity test does not seem to be known.

Finally, we would like to remark that, in the case $K = 1$, we can solve (15) and (20) explicitly, and we get Fieller's well-known confidence interval for the ratio of means, see Fieller (1940), Finney (1964).

3 Group Sequential Confidence Intervals on the Variance Parameter

Let $F_{\chi^2_\nu}$ denote the cumulative distribution function of a χ^2 -variate with ν degrees of freedom. Using the pivotal χ^2 -statistics from (2), which are monotone decreasing in $\sigma^2 > 0$, we obtain, in analogy to (5),

$$F_{\chi^2_{n_i - 2}} \left((n_i - 2) \frac{S_i^2}{\sigma^2} \right) \sim U(0, 1), \quad n_i = n_{E_i} + n_{C_i}, \quad i = 1, \dots, K,$$

leading to the combining pivotal statistics

$$Z_j^V(\sigma^2) = \sum_{i=1}^j \Phi^{-1} \left[F_{\chi_{n_i-2}^2} \left((n_i - 2) \frac{S_i^2}{\sigma^2} \right) \right] \sim \sqrt{j} \mathcal{N}(0, 1), \quad j = 1, \dots, K, \quad (25)$$

which are monotone decreasing in $\sigma^2 > 0$.

Since the confidence level for the variance parameter is usually chosen lower than for the outcome measure, let $cv_1(\kappa), \dots, cv_K(\kappa)$ denote critical values that yield the (one-sided) confidence level $1 - \kappa$, $0 < \kappa < 1/2$, in (8).

Let $\sigma_L^2(j)$ and $\sigma_U^2(j)$ be the unique solutions for σ^2 of the equations

$$Z_j^V(\sigma^2) = cv_j(\kappa) \text{ and } Z_j^V(\sigma^2) = -cv_j(\kappa), \quad j = 1, \dots, K. \quad (26)$$

Then, in analogy to (22), we build the confidence intervals for σ^2 as

$$VCI_k(\sigma^2) = [\sigma_{Lk}^2, \sigma_{Uk}^2], \quad (27)$$

where $\sigma_{Lk}^2 = \max\{\sigma_L^2(1), \dots, \sigma_L^2(k)\}$ and $\sigma_{Uk}^2 = \min\{\sigma_U^2(1), \dots, \sigma_U^2(k)\}$, $k = 1, \dots, K$.

The confidence intervals on σ^2 from (27) are nested, that is, $VCI_{k+1} \subset VCI_k$, $k = 1, \dots, K - 1$, and the confidence coefficients are at least $1 - 2\kappa$, $0 < \kappa < 1/2$. For small κ , an empty interval indicates that the assumptions of homogeneous variances over the stages may be violated.

Since descriptions of the standard deviation are preferable in documentation, we simply take the square root of the boundaries in VCI_k and denote the resulting confidence interval on σ by $VCI_k^{1/2}$.

4 Adaptive Sample Size Planning for the Effect Measure Ratio of Means

Planning with equal sample sizes in both two groups and suppressing the subscript i , we set $n_E = n_C = m$, $\xi = \mu_E - (1 - \Delta) \mu_C$, and $X = X_E - (1 - \Delta) X_C$, for a fixed value $\Delta \in [0, \Delta_0]$. Then we have with $\sigma(X)^2 = [1 + (1 - \Delta)^2] \sigma^2$, see (3),

$$X \sim \mathcal{N}(\xi, \sigma(X)^2) \text{ and } \bar{X} \sim \mathcal{N}\left(\xi, \frac{1}{m} \sigma(X)^2\right).$$

Let us first consider the test of the point hypotheses

$$H_0^* : \xi = 0 \quad \text{versus} \quad H_1^* : \xi = \xi^* > 0$$

using

$$T_0(1 - \Delta) = \sqrt{m} \frac{\bar{X}}{\sigma(X)} \sim \mathcal{N}(0, 1) \text{ under } H_0^*.$$

Then, for given level α , $0 < \alpha < 1$, and power $1 - \beta$, $0 < \beta < 1$, the required sample size m is given as (one-sample formula)

$$m = \frac{[\max\{0, \Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)\}]^2}{[\xi^*/\sigma(X)]^2}. \quad (28)$$

Let $s_0^2 > 0$ be a prior value for σ^2 and $\xi^* = \mu_{E_0} - (1 - \Delta)\mu_{C_0} > 0$ be a parameter value in the alternative $H_{1,\Delta}$, see (1), then the expected effect size $\xi^*/\sigma(X)$ in (28) is given as

$$A_0(\Delta) := \frac{\mu_{E_0} - (1 - \Delta)\mu_{C_0}}{s_0 \sqrt{1 + (1 - \Delta)^2}}. \quad (29)$$

Note that, for ease of presentation, we will use the normal sample size spending function from (28) in the following though we use t -statistics in fact. In practice, by correcting the sample size m with the variance of a t_{m-1} -variate, that is, by replacing m through $m(m-1)/(m-3)$, $m \geq 4$, the nearly exact sample size can be determined in a one-sample setting.

We introduce now two steering parameters for each stage j , say v_j and w_j , in order to cover a wide range of reasonable updating possibilities. We plan with equal sample sizes for both groups at each stage. After stage j , let be $A_j(\Delta) > 0$ a value for the effect size $\xi^*/\sigma(X)$, where $A_j(\Delta)$ is defined below. Then, conditioned on $A_j(\Delta) > 0$, a power of $1 - \beta$ at $\mu_E - (1 - \Delta)\mu_C = A_j(\Delta) \sigma \sqrt{1 + (1 - \Delta)^2}$ in the alternative $H_{1,\Delta}$ of (1) is approximately reached using t -statistics when the total sample size for both groups is chosen as

$$f_j(\alpha, \beta, \Delta) = 2 \frac{[\max\{0, \Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)\}]^2}{A_j(\Delta)^2}, \quad j = 0, 1, \dots, K, \quad (30)$$

with

$$\begin{aligned} A_j(\Delta) &= w_j \sum_{i=1}^j \frac{n_i}{\sum_{h=1}^j n_h} \frac{\bar{x}_{E_i} - (1 - \Delta) \bar{x}_{C_i}}{s_i \sqrt{1 + (1 - \Delta)^2}} \\ &\quad + (1 - w_j) \frac{\mu_{E_0} - (1 - \Delta)\mu_{C_0}}{[v_j s(j) + (1 - v_j) s_0] \sqrt{1 + (1 - \Delta)^2}} > 0, \\ s(j) &= \left(\sum_{i=1}^j \frac{n_i - 2}{\sum_{h=1}^j n_h - 2j} s_i^2 \right)^{1/2}, \quad n_i = n_{E_i} + n_{C_i}, \quad 0 \leq \Delta < 1, \\ &\quad 0 \leq w_j \leq 1, \quad w_0 = 0, \quad \text{and} \quad 0 \leq v_j \leq 1, \quad v_0 = 0, \end{aligned}$$

where \bar{x}_{E_i} , \bar{x}_{C_i} , and s_i^2 are the realizations of \bar{X}_{E_i} , \bar{X}_{C_i} , and S_i^2 up to stage j , $i = 1, \dots, j$.

Choosing $w_j = 0$ and $v_j = 0$ in (30), we obtain a purely prior information based sample size plan. The choice of $w_j = 0$ and $v_j > 0$ leads to adaptive plans that only use updated variances, where $s(j)^2$ denotes the pooled estimator of σ^2 up to stage j . For $w_j = 1$, the term $A_j(\Delta)$ is a version of the meta-analytical combination of effect sizes. Putting $w_j = 0$, when the first sample based estimate in $A_j(\Delta)$ is below the second one, gives priority to the second term as a lower bound. The reverse choice of w_j covers a situation considered in a two-stage trial by Liu and Chi (2001), and Proschan, Liu, and Hunsberger (2003), who also discussed the role of the effect size in updating sample sizes.

Recall now from (8) that it holds

$$\left\{ \sum_{i=1}^h Y_i \leq cv_h \text{ for } h = 1, \dots, j-1 \text{ and } \sum_{i=1}^{j-1} Y_i + \sqrt{K - (j-1)} Y_j \leq cv_K \right\} \quad (31)$$

$$\supset \left\{ \sum_{i=1}^h Y_i \leq cv_h \text{ for all } h = 1, \dots, K \right\}.$$

In the group sequential trial, the hypothesis $H_{0,\Delta}$ will be rejected if $Z_j(1 - \Delta) > cv_j$, see (7) and (9). Then, by (31), if we decide after stage $j - 1$ to omit the interim analyses j up to $K - 1$, we can assign the remaining weight $\sqrt{K - j + 1}$ to the next final study part and build the final test statistic as

$$Z_{j,K}(1 - \Delta) = Z_{j-1}(1 - \Delta) + \sqrt{K - j + 1} \Phi^{-1} [F_{t_{n_{j-2}}}(T_j(1 - \Delta))]. \quad (32)$$

Note that it holds $Z_{j,K}(1 - \Delta) \sim \sqrt{K} \mathcal{N}(0, 1)$ under $H_{0,\Delta}$, $j = 1, \dots, K$, and define $Z_0(1 - \Delta) = 0$. The test statistic $Z_{j,K}(1 - \Delta)$ has to be compared with the K -th critical value cv_K in testing $H_{0,\Delta}$.

The p -value of testing $H_{0,\Delta}$ is given at stage i by use of $T_i(1 - \Delta)$ as

$$p_i = p_i(\Delta) = 1 - F_{t_{n_i-2}}(T_i(1 - \Delta)), \quad i = 1, \dots, K. \quad (33)$$

Provided that a significant result has not been reached in the first $j - 1$ stages, that is, $Z_i(1 - \Delta) \leq cv_i$ for $i = 1, \dots, j - 1$, and in the next study part we want to attain cv_K by use of the final test statistic

$$\hat{Z}_{j,K}(1 - \Delta) = Z_{j-1}(1 - \Delta) + \sqrt{K - j + 1} \Phi^{-1}(1 - \hat{p}_{j,K}(\Delta)),$$

then the projected p -value $\hat{p}_{j,K}(\Delta)$ of the next study part should be

$$\hat{p}_{j,K}(\Delta) = 1 - \Phi \left[(cv_K - Z_{j-1}(1 - \Delta)) / \sqrt{K - j + 1} \right]. \quad (34)$$

To detect a deviation of the null-hypothesis $H_{0,\Delta}$ in direction of H_1,Δ at $\mu_E - (1 - \Delta)\mu_C = A_{j-1}(\Delta) \sigma \sqrt{1 + (1 - \Delta)^2} > 0$ with the (conditional) power $(1 - \beta)$, the sample size for the next final study stage must be chosen using (30) as

$$M_j = M_j(\Delta) = f_{j-1}(\hat{p}_{j,K}(\Delta), \beta, \Delta). \quad (35)$$

If we do not want to finish the trial in this way and have in mind the originally planned $K - (j - 1)$ further stages, we suggest choosing the sample size of stage j proportionally as

$$n_j = n_j(\Delta) = \frac{M_j(\Delta)}{K - j + 1}, \quad n_{E_j} = n_{C_j} \approx n_j/2, \quad j = 1, \dots, K. \quad (36)$$

Recall that the sample sizes should be at least 2 in each stage and group.

For $j = 1$ and $\hat{p}_{1,K} = 1 - \Phi\left(cv_K/\sqrt{K}\right)$, we obtain the starting sample size of the group sequential trial as

$$n_1 = M_1/K, \quad (37)$$

where

$$M_1 = 2 \left(\frac{cv_K}{\sqrt{K}} + \Phi^{-1}(1 - \beta) \right)^2 / A_0(\Delta)^2,$$

with $A_0(\Delta) > 0$ from (29) and $0 < \beta < 1/2$. Since $A_0(1 - \Delta)$ has the same structure as $T_i(\Delta)$, $A_0(\Delta)$ is increasing in Δ , see Lemma 2, and n_1 is maximum for $\Delta = 0$, that is, the sample size is maximum for the superiority trial as it should be.

Starting with n_1 observations from (37) in the first stage and applying the proceeding presented above, we reach the full power $1 - \beta$, conditioned on

$$\mu_E - (1 - \Delta)\mu_C = A_{K-1}(\Delta) \sigma \sqrt{1 + (1 - \Delta)^2} > 0,$$

latest in stage $j = K$, if the trial does not stop earlier because of shown significance.

Note that, see (5) and (33), we may formally define the p -values as suiting to the null-hypothesis that λ is the *true* parameter, see Cox and Hinkley (1974, p. 221). So we may apply the general result that, under the null-hypothesis, p -values preserve their distribution and independence (for continuous null-distributions) when sample sizes are adaptively chosen in a consecutive way, see Brannath, Posch, and Bauer (2002). All the presented procedures are based on such p -values. Consequently, the statements of Section 2 and 3 remain valid when sample sizes are adaptively chosen as demonstrated in this section, see also Hartung (2006).

Table 1: Data and (≥ 0.95)-confidence intervals on λ and σ using a noninferiority margin $\Delta_0 = 10\%$, and prior guess in stage 0.

| Stage | Sample size | Data (in ℓ) on | | | Confidence intervals on | |
|-------|-------------|----------------------|---------|----------|---------------------------|------------------|
| i | n_i | μ_E | μ_C | σ | $\lambda = \mu_E/\mu_C$ | σ |
| 0 | - | 2.75 | 2.50 | 0.75 | [$1 - \Delta_0 = 0.90$] | |
| 1 | 128 | 2.67 | 2.55 | 0.81 | [0.8604, 1.2765] | [0.6621, 1.0287] |
| 2/3 | 56 | 2.70 | 2.56 | 0.87 | [0.9483, 1.1646] | [0.7547, 0.9482] |

5 A Real Data Example

Let us consider a clinical trial one of the authors was concerned with as a statistical advisor. Two different inhalers, a new drug (E) and a standard drug (C), for treating patients with asthma bronchiale are compared with respect to a lung function parameter named FEV₁: forced expiratory volume in 1 second, measured in liter (ℓ). The ratio of means is the common outcome measure and an often used noninferiority margin is $\Delta_0 = 0.10$ in this field of application. The one-sided significance level is chosen as $\alpha = 0.025$ and the power should be $1 - \beta = 0.90$ at the expected effect size. The sample sizes of the two treatment groups are equal at each stage and the drugs are equally randomized within blocks of size 8. The involved investigators were optimistic so that the trial started with an attempt to show superiority. That means $\Delta = 0$ in the testing problem (1), which then is equivalent to testing the difference of the means. So we could use sample size calculations of usual group sequential trials for the purpose of comparison, for instance, from Jennison and Turnbull (2000).

Since no early significant results were expected, the decision was made in favor of an adaptive three-stage design of O'Brien and Fleming (1979) type. Using the combining statistic (7), we obtain the constant critical values $cv_j = cons_{OBF}(3, 0.025) = 3.471 (= 2.004 \sqrt{3})$ for $j = 1, 2, 3$ in (8), see Jennison and Turnbull (2000, p. 29), Hartung (2006, p. 533). Planning with $\Delta = 0$ for showing superiority, we obtain the prior guess $A_0(0) = 0.2357$, see (29), using the prior guesses for the means and the standard deviation from Table 1.

Using (30), we would need 378 patients in a one-stage fixed sample size test. With Jennison and Turnbull (2000, p. 30), each stage in a three-stage fixed sample size O'Brien and Fleming test should have $1.016 \times 378/3 \approx 128$ patients for testing $H_{0,0} : \mu_E = \mu_C$

versus $H_{1,0} : \mu_E > \mu_C$ and σ^2 known. Formula (37) yields the value $388/3 \approx 130$ patients as starting sample size n_1 .

We started with $n_1 = 128$ patients because of the randomization scheme using blocks of size 8. In the first stage, we observed $\bar{x}_{E_1} = 2.67$, $\bar{x}_{C_1} = 2.55$, and $s_1 = 0.81$, see Table 1, that led to $Z_1(1) = 0.8352$, see (7). Equating $Z_1(\lambda)$ to the critical values 3.471 and to -3.471 and solving for λ yields the first confidence interval on the ratio $\lambda = \mu_E/\mu_C$ as $CI_1 = [0.8604, 1.2765]$.

It turned out that the prior guesses of the parameters were too optimistic with respect to the new drug. Therefore, we used only the observed values from the first stage in the further planning, especially as these values were based on a relatively large number of patients. Consequently, the weight w_j in the sample size spending function (30) was set to $w_j = 1$ for the next stage leading to $A_1(0) = 0.1048$. The projected p -value according to (34) is

$$\hat{p}_2(0) = 1 - \Phi\left([3.471 - 0.835]/\sqrt{2}\right) = 0.0312,$$

leading to $n_2(0) \approx 902 = 1804/2$ patients for the next stage for showing superiority using (36). Since the interest in showing superiority came from a marketing point of view, the easy decision was made to continue the trial with showing noninferiority being sufficient for regulatory concerns.

Consequently, we used $\Delta = \Delta_0 = 0.10$ in the further planning. By (7), we compute $Z_1(0.90) = 2.7075$ and, by (34), the projected p -value

$$\hat{p}_2(0.10) = 1 - \Phi([3.471 - 2.7075]/\sqrt{2}) = 0.2946.$$

In the sample size spending function (30), we obtain now $A_1(0.10) = 0.3441$ with $w_1 = 1$. So, formula (35) yields $M_2(0.10) \approx 56$ patients for the total sample size of the two remaining planned stages leading to 32 patients in the second stage because of the block size of 8. Therefore, it was decided to save the effort with an additional interim analysis and to omit the originally planned second one. In the final study part, say stage 2/3, the sample size is $n_{2/3} = 56$.

If we would use now the combining statistic (7), we could lose potential power of the test since we would not make full use of the predefined level α . However, based on the considerations in Section 4, we can use the combining statistic

$$Z_{2,3}(\lambda) = Z_1(\lambda) + \sqrt{2} \Phi^{-1} \left[F_{t_{n_{2/3}-2}}(T_2(\lambda)) \right]$$

in the final analysis, see (32). The final test statistic has then to be compared with the critical value cv_3 .

In the final stage 2/3, we observed $\bar{x}_{E_2} = 2.70$, $\bar{x}_{C_2} = 2.56$, and $s_2 = 0.87$, see Table 1. Equating $Z_{2,3}(\lambda)$ to 3.471 and to -3.471 and solving for λ yielded the second confidence interval on the ratio λ as $CI_{2/3} = [0.9483, 1.1646]$, which lies clearly above 0.90.

Further we calculated the final test value,

$$Z_{2,3}(0.90) = 2.7075 + \sqrt{2} \times 1.7564 = 5.1914 > 3.471,$$

which confirms significant noninferiority of the new drug with regard to the standard drug.

For the confidence intervals on the variance parameter, we chose the same level $\kappa = 0.025$, see Section 3, so that we can use the same critical values as above. In the first stage, we had to equate $Z_1^V(\sigma^2) = \Phi^{-1} \left[F_{\chi_{126}^2} (126 \times 0.81^2 / \sigma^2) \right]$ to 3.471 and to -3.471 and solving for σ^2 yielded the first confidence interval on σ^2 as $VCI_1(\sigma^2) = [0.4384, 1.0582]$.

In the final stage, we use, in analogy to the considerations above, the combining statistic $Z_{2,3}^V(\sigma^2) = Z_1^V(\sigma^2) + \sqrt{2} \Phi^{-1} \left[F_{\chi_{54}^2} (54 \times 0.87^2 / \sigma^2) \right]$. Equating $Z_{2,3}^V(\sigma^2)$ to 3.471 and to -3.471 and solving for σ^2 yielded the second confidence interval on σ^2 as $VCI_{2/3}(\sigma^2) = [0.5696, 0.8991]$. Taking the square root of the σ^2 -boundaries, we get the confidence intervals on σ as presented in Table 1.

6 Final Remarks

In this paper, we defined a two-sided confidence interval CI_k on the ratio of normal means by intersecting the lower and upper one-sided confidence intervals $CI_{k,L}$ and $CI_{k,U}$ on this parameter with confidence coefficients of $1 - \alpha$, respectively. The confidence coefficient of the two-sided interval CI_k is always at least $1 - 2\alpha$. If we use the critical values of the correspondent two-sided test at level 2α , we will get a two-sided confidence interval, say CI_k^0 , that is slightly narrower than CI_k for $K \geq 2$, but has a confidence coefficient of at least $1 - 2\alpha$ as well. Moreover, the final interval CI_K^0 has a confidence coefficient of exactly $1 - 2\alpha$. However, using the lower boundary of CI_k^0 for the test decision in the test problem (1) as given at the end of Section 2.1, the test level α cannot be guaranteed. However, no severe differences between both two-sided intervals and test decisions are expected for practical applications, because, at least for $\alpha \leq 0.05$, the involved critical values are nearly identical, see Jennison and Turnbull (2000, p. 192).

In Section 2.2, we presented the two-sided confidence intervals as $CI_k = CI_{k-1} \cap I_k$, see Remark 1, where I_k results from the boundaries in stage k alone and, thus, is always

nonempty. Since I_k has also a confidence coefficient of at least $1 - 2\alpha$, we could prefer I_k to CI_k , see, for instance, Jennison and Turnbull (2000, p. 192) in their corresponding settings. But, suppose that up to stage $k - 1$ the intersections CI_ℓ , $\ell = 1, \dots, k - 1$, are nonempty and CI_k is empty for a common level α . That means, the nonempty interval I_k lies completely outside the nonempty interval CI_{k-1} . Therefore, up to an error rate of 2α , see (24), we may consider that the ratios of the underlying mean treatment effects are no longer homogeneous in all the stages. So, with regard to statistical concerns, results from this stage k should not influence conclusions, or non-conclusions, from the previous stages. Consequently, preferring I_k to CI_k does not provide a real advantage. Nevertheless, one may try to find an explanation for the revealed treatment behavior. Recall that under the model assumption of an identical ratio μ_E/μ_C in all the stages of the trial, the probability obtaining an empty interval CI_k is bounded by 2α .

Moreover, let us consider the testing problem in a group sequential trial. In a superiority test, the null-hypothesis $H_{0,0}$ is rejected at level α in favor of $H_{1,0}$ if we observe $Z_{k^*}(1) > cv_{k^*}$ in at least one stage $k^* \in \{1, \dots, K\}$ or, equivalently, the lower bound of I_{k^*} lies above 0. Suppose $k^* < K$ and the study is continued to obtain more data, for instance, for safety reasons in clinical trials. Then, in all further stages $k > k^*$, we may observe $Z_k(1) \leq cv_k$ or that I_k covers 0 without contradicting the already shown superiority. This fact is able to induce misunderstandings in practical applications caused by a lack of knowledge of the theoretical background. The same problem may arise when, after a shown significant noninferiority, the trial is continued for an attempt to show superiority. Such possible misunderstandings are avoided by using CI_k instead of I_k as proposed in the testing procedure, see Section 2.1. The automatically implied homogeneity test (24) would react if quite different results would have been observed in later stages.

Finally, let us briefly address point estimation of the parameters λ and σ^2 . The combined statistic $Z_j(\lambda)$ from (7) is a $\mathcal{N}(0, j)$ -distributed random variable with mode and median 0. A maximum likelihood (ML) estimator $\hat{\lambda}_{ML}(j)$ of the ratio $\lambda = \mu_E/\mu_C$ at stage j is given as the solution of $Z_j(\hat{\lambda}_{ML}(j)) = 0$. The global p -value at stage j is $p_G(j) = 1 - \Phi(Z_j(\lambda)/\sqrt{j})$, and solving the equation for $p_G(j) = 1/2$ yields $\hat{\lambda}_{ML}(j)$ as the solution. Note, that $Z_j(\lambda)$ is monotone in λ . Consequently, $\hat{\lambda}_{ML}(j)$ is a median unbiased estimator, see Cox and Hinkley (1974, p. 273).

For the variance parameter σ^2 , we use the combining statistic $Z_j^V(\sigma^2)$ from (25). The ML-estimator $\hat{\sigma}_{ML}^2(j)$ of σ^2 at stage j is given as the solution of $Z_j^V(\hat{\sigma}_{ML}^2(j)) = 0$. By the same considerations as above, $\hat{\sigma}_{ML}^2(j)$ is a median unbiased estimator.

References

- Bauer, P. and Kieser, M. (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika* **83**, 934–937.
- Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97**, 236–244.
- Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, New York.
- EMA (The European Agency for the Evaluation of Medicinal Products) (2002). *Points to consider on Switching between Superiority and Noninferiority*. London, CPMP/EWP/482/99.
- Fieller, E. C. (1940). The biological standardization of insulin. *Journal of the Royal Statistical Society* (Suppl.) **7**, 1-64.
- Finney, D. J. (1964). *Statistical Methods in Biological Assay*. 2nd Edition. Griffin, London.
- Hartung, J. (2006). Flexible designs by adaptive plans of generalized Pocock- and O'Brien-Fleming-type and by Self-designing clinical trials. *Biometrical Journal* **48**, 521–536.
- Hartung, J. and Knapp, G. (2006). Repeated confidence intervals in Self-designing clinical trials and switching between noninferiority und superiority. *Biometrical Journal* **48**, 697–709.
- Hartung, J., Knapp, G., and Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications*. Wiley, New York.
- Jennison, C. and Turnbull, B. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, Boca Raton and London.
- Liu, Q. and Chi, G. Y. H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics* **57**, 172–177.
- 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.

Proschan, M. A., Liu, Q., and Hunsberger, S. (2003). Practical midcourse sample size modification in clinical trials. *Controlled Clinical Trials* **24**, 4–15.