A Confidence Interval Approach for Difference and Ratio of Normal Means in Self-designing Clinical Trials

Joachim Hartung[1] **and Guido Knapp**

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

Abstract: In Self-designing clinical trials, confidence intervals are derived for the difference and the ratio of normal means, where the results of the independent study stages are combined using the weighted inverse normal method. The confidence intervals always hold the predefined nominal confidence level. During the course of the Self-designing trial, the sample sizes as well as the number of study stages can be determined simultaneously in a completely adaptive way. Self-designing may be considered as the limit case of adaptive group sequential designing of O'Brien and Fleming type when the full significance level is shifted to the last stage. We consider the effect measures difference and ratio of normal means, where the latter has not yet been considered in group sequential trials so far. Concrete rules are derived for updating sample sizes and assigning weights to the stages of the trial. The clinical trial may be originally designed either to show non-inferiority or superiority. But, in each interim analysis, it is possible to change the planning from showing non-inferiority to showing superiority or vice versa. The performance of the Self-designing and the resulting confidence intervals are demonstrated in real-data examples for both considered effect measures showing both kinds of switching during an ongoing trial.

Keywords: Adaptive planning; Confidence interval; Learning rule; Ratio of means; Self-designing; Switching between non-inferiority and superiority; Weighted inverse normal method.

1 Introduction

In a clinical examination, the common effect measures for comparing a new agent to a standard agent with regard to (at least) non-inferiority are the difference of means and the ratio of means. Provided the standard agent is well known and stable in different populations, the suitable measure is the difference of means. Otherwise, the scale invariant ratio

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

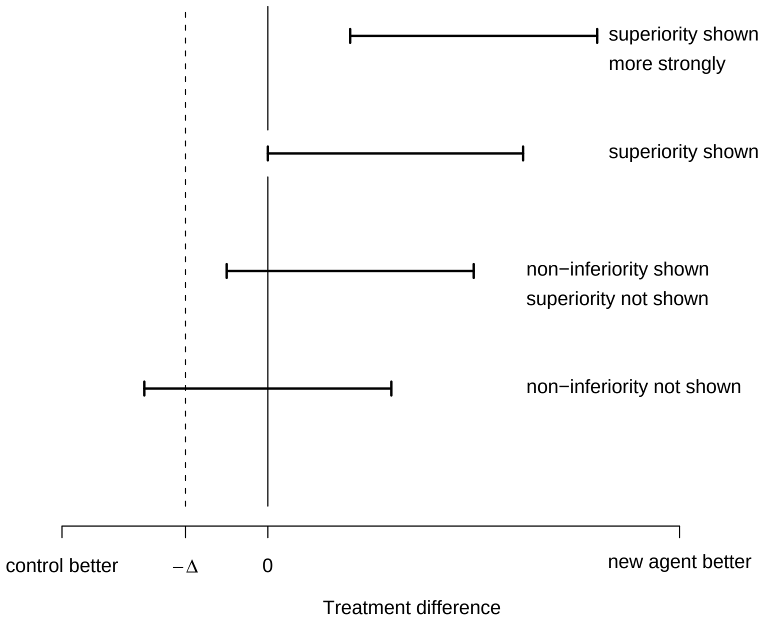

Figure 1: Examples of final 95%-confidence intervals for different study results.

of means is the preferred effect measure. In the analysis, the confidence interval approach is of particular attractiveness, see e. g. EMEA (2000). From that guideline we also take over the graphical illustration of switching from non-inferiority over to superiority, see Figure 1.

The theoretical background for switching between non-inferiority and superiority is discussed, for example, by Bauer and Kieser (1996) and Brannath et al. (2003). Practically this means that the position of the confidence interval determines the kind of result of the study, independently of the question whether originally the study was planned as non-inferiority or superiority trial.

In classical group sequential trials, the repeated confidence interval approach introduced by Jennison and Turnbull (1984, 1989) may be applied for constructing confidence intervals on the parameter of interest. For adaptive clinical trials, several proposals for constructing a confidence interval exist for various kinds of flexible designs, see, for instance, Lehmacher and Wassmer (1999), Liu and Chi (2001), Brannath, Posch, and Bauer (2002), Brannath, König, and Bauer (2003), Frick (2002), Proschan, Liu, and Hunsberger (2003), and Hartung and Knapp (2006).

In the following, we consider flexible adaptive group sequential trials in the sense that, besides the adaptive choice of the sample sizes for the different stages, the number of stages can be either fixed in advance or can be determined also in an adaptive way, the latter approach named Self-designing as introduced by Fisher (1998), Shen and Fisher (1999).

In the Self-designing approach of group sequential trials, one decides adaptively after each interim-analysis during the course of the study whether exactly one or at least two further study stages will be performed by use of the unblinded results of all the already conducted interim-analyses. The Self-designing trial ends when the (finite) variance of an a priori fixed final test statistic is used up. Hartung (2001) derives Self-designing rules where the weighted inverse normal method is used for combining the p-values of the independent study stages. Simultaneously the weights and the sample sizes can be chosen adaptively. Considering the adaptive extension of O'Brien and Fleming (1979) designs, Self-designing can be viewed as the limit case when the needed level attained of the last stage reaches the full overall significance level, see Hartung (2006). It should be mentioned, that in spite of its practical importance, the effect measure ratio of means is not considered in group sequential trials until now.

In a Self-designing trial, Cheng and Shen (2004) construct a confidence interval for the mean difference of two normal variates, where the variance parameter is assumed to be known. As in Shen and Fisher (1999), the sequence of possible sample sizes is fixed in advance and just the weights assigned to the stages of the trial are really chosen adaptively. For unknown variance, Cheng and Shen (2004) give an approximate solution.

Extending the proceeding of Hartung (2001, 2006) to the combination of parameterized p-values, we will derive exact confidence intervals for both effect measures, difference and ratio of normal means, with unknown variance parameter. Moreover, a confidence interval for the variance parameter will also be derived. For both effect measures, suitably combined learning rules provide an effective chance to choose both sample sizes and weights simultaneously in an adaptive way. In our approach, we consider t-statistics involving the unknown parameter and combine them using the weighted inverse normal method from meta-analysis, see Hedges and Olkin (1985) or Hartung, Knapp, and Sinha (2008). The confidence intervals are defined implicitly and, for the determination of the boundaries, nonlinear equations have to be solved, whose solutions are unique.

In each interim analysis we may decide in the planning between non-inferiority and superiority. Based on conditional error functions, we derive concrete rules for adaptive designing, ranging from fixed prior information based planning over just updating of variances up to completely data based planning. Our proceeding is a conditional power approach, as applied at least implicitly, for instance, by Proschan and Hunsberger (1995), Denne (2001), Liu and Chi (2001), Proschan, Liu, and Hunsberger (2003) in two-stage adaptive designs, and by Shen and Fisher (1999), Hartung (2000, 2001, 2006), Hartung and Knapp (2003, 2006), Cheng and Shen (2004) in the context of Self-designing clinical trials.

The outline of the paper is as follows. In Section 2, the basics for a Self-designing study of comparing normal outcomes are summarized. The construction of a confidence interval for the mean difference is described in Section 3. Section 4 contains the adaptive planning for sample sizes and weights when the mean difference is the parameter of interest. Moreover, the switching of the planning between non-inferiority and superiority is addressed. The construction of a confidence interval for the variance parameter is discussed in Section 5, and in Section 6, an example is considered in which the methods presented so far are illustrated. Section 7 contains the construction of a confidence interval when the ratio of normal means is the parameter of interest. Moreover, some considerations of adaptive planning in this situation are discussed. In Section 8, the methods of the previous section are illustrated in an example. Finally, some concluding remarks are given, where also point estimation of the considered effect measures is addressed.

2 Basic principles for a Self-designing study of comparing normal outcomes

Let x_E and x_C be independent normally distributed random variables with mean μ_E in an experimental group E and mean μ_C in an (active) control group C with common variance $\sigma^2 > 0$, that is, succinctly

$$x_E \sim \mathcal{N}(\mu_E, \sigma^2) \quad \text{and} \quad x_C \sim \mathcal{N}(\mu_C, \sigma^2)\,. \tag{1}$$

A comparative study is carried out consecutively in a number of, say k, independent stages, denoted by $stg(1)$, ..., $stg(k)$. In the i-th stage, $i = 1, \dots, k$, let us observe the responses x_{Eij}, $j = 1, \dots, n_{Ei} \geq 2$, and x_{Cij}, $j = 1, \dots, n_{Ci} \geq 2$, where n_{Ei} and n_{Ci} are the sample sizes in the respective groups. The observed mean difference measure in $stg(i)$ is

$$y_i = \frac{1}{n_{Ei}} \sum_{j=1}^{n_{Ei}} x_{Eij} - \frac{1}{n_{Ci}} \sum_{j=1}^{n_{Ci}} x_{Cij} = \bar{x}_{Ei} - \bar{x}_{Ci}, \quad i = 1, \dots, k. \tag{2}$$

The variance parameter σ^2 is estimated in the i-th stage by the pooled estimator

$$s_i^2 = \frac{1}{n_{Ei} + n_{Ci} - 2} \left(\sum_{j=1}^{n_{Ei}} (x_{Eij} - \bar{x}_{Ei})^2 + \sum_{j=1}^{n_{Ci}} (x_{Cij} - \bar{x}_{Ci})^2 \right), \quad i = 1, \ldots, k, \tag{3}$$

which follows a scaled χ^2-distribution with $n_{Ei} + n_{Ci} - 2$ degrees of freedom, that is,

$$(n_{Ei} + n_{Ci} - 2) \, \frac{s_i^2}{\sigma^2} \; \sim \; \chi^2(n_{Ei} + n_{Ci} - 2). \tag{4}$$

The variance of y_i is estimated in the i-th stage by

$$\widehat{\mathrm{var}}(y_i) = \left(\frac{1}{n_{Ei}} + \frac{1}{n_{Ci}} \right) \, s_i^2, \tag{5}$$

and y_i and s_i^2 are stochastically independent, $i = 1, \ldots, k$.

Let us assign a positive normed weight $w_i > 0$ to each stage i, $i = 1, \ldots, k$, with $\sum_{i=1}^{k} w_i = 1$. Based on considerations in Fisher (1998), Shen and Fisher (1999), Hartung (2001, 2006), and Cheng and Shen (2004), the sample sizes as well as the weights may be chosen in a completely adaptive way. All the information of the unblinded data of previous stages can be used to choose simultaneously the sample size and the weight for the next stage. Let $stg(0)$ denote a priori information and external restrictions, we express the adaptive choice of sample sizes and weights as

$$n_i = \hat{n}\{i-1\} = \hat{n}\{stg(0), stg(1), \ldots, stg(i-1)\}, \; n_i = n_{Ei} + n_{Ci}, \tag{6}$$

and

$$w_i = \hat{w}\{i-1\} = \hat{w}\{stg(0), stg(1), \ldots, stg(i-1)\}, \tag{7}$$

where $w_i \leq 1 - w_\Sigma(i-1)$, $w_\Sigma(i) = \sum_{j=1}^{i} w_i$, $w_\Sigma(0) = 0$, $w_\Sigma(k) = 1$, $w_i > 0$, $i = 1, \ldots, k$.

Note that the number k of performed stages is random and will be realized during the course of the sequential trial in dependence of the choice of weights. Of course, k has to be finite (almost surely), and for practical reasons, k should be bounded by some reasonable constant. Introducing a minimum weight, say $w_{\min}$, $0 < w_{\min} < 1$, for a realized stage, we obtain the boundary as $k \leq 1/w_{\min}$. A minimum sample size, say $n_{\min}$, may also be introduced, so that

$$n_i \geq n_{\min} \geq 4 \qquad \text{and} \qquad w_i \geq w_{\min} > 0, \quad i = 1, \ldots, k. \tag{8}$$

The use of minimum weight and minimum sample size leads to useful termination conditions of the whole trial and can adjust some non-practicable suggestions of the (automatic) learning rules for choosing n_i and w_i discussed in later sections.

3 A confidence interval for the mean difference

With an a priori defined non-inferiority bound $\Delta_0 \geq 0$, we are interested in testing

$$\mathrm{H}_{0,\Delta}: \ \mu_E \leq \mu_C - \Delta \quad \text{versus} \quad \mathrm{H}_{1,\Delta}: \ \mu_E > \mu_C - \Delta \ , \quad 0 \leq \Delta \leq \Delta_0, \tag{9}$$

at a prescribed level α, $0 < \alpha < 1/2$. The alternative hypothesis $\mathrm{H}_{1,\Delta}$ means ($\Delta-$)non-inferiority for $0 < \Delta \leq \Delta_0$, and, for $\Delta = 0$, superiority of E with regard to C.

Let $\vartheta = \mu_E - \mu_C$ denote the difference of means, which can be unbiasedly estimated by y_i in $stg(i)$, $i = 1, \ldots, k$, see (2). For the i-th stage, let us define the t-statistic

$$T_i(\vartheta) = \frac{y_i - \vartheta}{\sqrt{(1/n_{Ei} + 1/n_{Ci})\ s_i^2}} \sim t(n_{Ei} + n_{Ci} - 2)\ , \tag{10}$$

that is, for the true parameter ϑ, the statistic $T_i(\vartheta)$ follows a (central) t-distribution with $n_{Ei} + n_{Ci} - 2$ degrees of freedom.

Let $F_{t(\nu)}$ denote the cumulative distribution function of a t-variate with ν degrees of freedom, then it holds, for the $1-p$-value,

$$F_{t(n_{Ei}+n_{Ci}-2)}(T_i(\vartheta)) \sim \mathcal{U}(0,1), \quad i = 1, \ldots, k, \tag{11}$$

where $\mathcal{U}(0,1)$ stands for the uniform distribution in the unit interval. Then, we have

$$z_i(\vartheta) = \Phi^{-1}[F_{t(n_{Ei}+n_{Ci}-2)}(T_i(\vartheta))] \sim \mathcal{N}(0,1)\ ,\ i = 1, \ldots, k, \tag{12}$$

with Φ^{-1} the inverse of the standard normal distribution function Φ. Although sample sizes and weights may be chosen adaptively as described in (6) and (7), the final combining statistic follows a specified test distribution, that is,

$$Z_k(\vartheta) = \sum_{i=1}^{k} \sqrt{w_i}\, z_i(\vartheta) \sim \mathcal{N}(0,1)\ , \quad \text{with} \quad w_\Sigma(k) = \sum_{i=1}^{k} w_i = 1, \tag{13}$$

see Fisher (1998), Shen and Fisher (1999), and Hartung (2001).

The continuous distribution functions $F_{t(\nu_i)}(\cdot)$ and the inverse distribution function $\Phi^{-1}(\cdot)$ are (strictly) monotone increasing functions in their arguments. The pivotal statistic $T_i(\vartheta)$ from (10) is monotone decreasing in ϑ, implying that $\Phi^{-1}(F_{t(\nu_i)}(T_i(\vartheta))$ is monotone decreasing in ϑ. Hence, the whole function $Z_k(\vartheta)$ is monotone decreasing in ϑ.

So we can define the following confidence interval on ϑ,

$$\mathrm{CI}(\vartheta) = \left\{ d \in I\!R \mid \Phi^{-1}(\alpha) \leq Z_k(d) \leq \Phi^{-1}(1-\alpha) \right\} = [\ \vartheta_L\ ,\ \vartheta_U\ ] \tag{14}$$

where ϑ_L and ϑ_U are the unique solutions of the equations:

$$Z_k(\vartheta_L) = \Phi^{-1}(1-\alpha) \quad \text{and} \quad Z_k(\vartheta_U) = -\Phi^{-1}(1-\alpha).$$

The confidence coefficient of $\text{CI}(\vartheta)$ is $1 - 2\alpha$, $0 < \alpha < 1/2$. Since the solutions in (14) are unique, they can easily be found iteratively using standard statistics software packages.

Let us now apply the confidence interval to the test problem (9). We decide, at level α, for the alternative $\text{H}_{1,\Delta}$, $\Delta \in [0\,,\,\Delta_0]$, if $-\Delta$ lies below $\text{CI}(\vartheta)$, and we do not reject H_{0,Δ_0}, if $\text{CI}(\vartheta)$ covers $-\Delta_0$, more succinctly, with ϑ_L from (14),

$$\begin{aligned} &\text{if } -\Delta < \vartheta_L\,, \quad &&\text{then reject } \text{H}_{0,\Delta}, \\ &\text{if } -\Delta_0 \geq \vartheta_L\,, \quad &&\text{then stay with } \text{H}_{0,\Delta_0}. \end{aligned} \tag{15}$$

Let us briefly consider the case that the variance parameter is known in advance, say σ_0^2. Then the statistic (10) becomes the z-statistic

$$T_{i,0}(\vartheta) = \frac{y_i - \vartheta}{\sqrt{1/n_{Ei} + 1/n_{Ci}}\;\sigma_0} = \frac{y_i - \vartheta}{\sigma(y_i)} \sim \mathcal{N}(0,1). \tag{16}$$

With $z_i(\vartheta) = \Phi^{-1}(\Phi(T_{i,0}(\vartheta))) = T_{i,0}(\vartheta)$, $Z_k(\vartheta)$ in (13) becomes $Z_{k,0}(\vartheta) = \sum_{i=1}^{k} \sqrt{w_i} T_{i,0}(\vartheta) \sim \mathcal{N}(0,1)$. Equating now $Z_{k,0}(\vartheta) = \pm\Phi^{-1}(1-\alpha)$ and solving for ϑ yields the $(1-2\alpha)$-confidence interval on ϑ

$$\text{CI}_0(\vartheta) = \left[\sum_{i=1}^{k} \frac{\sqrt{w_i} y_i / \sigma(y_i)}{\sum_{h=1}^{k} \sqrt{w_h}/\sigma(y_h)} \;\pm\; \frac{\Phi^{-1}(1-\alpha)}{\sum_{h=1}^{k} \sqrt{w_h}/\sigma(y_h)} \right]. \tag{17}$$

This interval is also considered, in a different presentation, by Cheng and Shen (2004). Replacing σ_0^2 by the observed values s_i^2 leads to approximate z-statistics in (16) and an approximate confidence interval in (17). Note that the combined test statistics of Fisher (1998) and Shen and Fisher (1999) are also special cases of the general weighted inverse normal combining statistics, see Hartung (2006).

4 Adaptive planning for sample sizes and weights

The confidence interval $\text{CI}(\vartheta)$ in (14) results after $k-1$ interim analyses based on the unblinded data. In case an unexpected favorable parameter constellation has been observed up to stage j and provided that $w_\Sigma(j) < 1$, this may lead to considerations to switch from showing non-inferiority to showing superiority, and so the trial is then continued by further planning with $\Delta = 0$. Conversely, originally planned as a superiority trial, a first interim analysis may reveal that an unexpected large number of subjects would be required. So, in case of an active control, one may decide to switch from showing superiority to showing non-inferiority, 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 non-inferiority bound Δ_0 should have been defined at the beginning of the study, see also the discussion in the guideline EMEA (2000). In the following, we present some learning rules for choosing the sample sizes and the weights adaptively with the possibility of switching in the planning between non-inferiority and superiority. Moreover, we chose two real-data examples to demonstrate that both kinds of switching may occur during ongoing trials in a quite natural way, see Sections 6 and 8.

For predefined type I and II error rates α, $0 < \alpha < 1$, and β, $0 < \beta < 1$, respectively, let us consider, for ease of presentation, the approximate normal sample size spending function. Two steering parameters u_j and v_j will be introduced for each stage j in order to cover a wide range of reasonable updating possibilities, whose realization would then depend on a given concrete situation. We plan with equal sample sizes for both groups at each stage. Based on information up to stage j, an estimate $A_j(\Delta) > 0$ of the standardized mean difference $(\vartheta + \Delta)/\sigma$ may be assumed, where $A_j(\Delta)$ is defined below. The power is considered at the point $\vartheta + \Delta = \sigma A_j(\Delta)$ in the alternative $H_{1,\Delta}$. For testing $H_{0,\Delta}$ from (9) by use of a t-test of level α at stage $j+1$, a power of $1-\beta$ is approximately reached when the total sample size for both groups at stage $j+1$ is chosen as

$$f_j(\alpha, \beta, \Delta) = \frac{4\,[\max\{0\,,\ \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}]^2}{A_j(\Delta)^2}, \qquad j = 0, 1, \ldots, k, \tag{18}$$

with

$$A_j(\Delta) = u_j \sum_{i=1}^{j} \frac{\tilde{n}_i}{\sum_{h=1}^{j} \tilde{n}_h} \left(\frac{y_i + \Delta}{s_i} \right) + (1-u_j) \frac{\mu_{E0} - \mu_{C0} + \Delta}{v_j\ s(j) + (1-v_j)\ s_0} > 0, \quad \Delta \geq 0,$$

$$s(j) = \left(\sum_{i=1}^{j} \frac{n_i - 2}{\sum_{h=1}^{j} n_h - 2j}\ s_i^2 \right)^{1/2}, \quad \tilde{n}_i = \frac{2}{1/n_{E_i} + 1/n_{C_i}},$$

$$n_i = n_{E_i} + n_{C_i}, \qquad 0 \leq u_j \leq 1, u_0 = 0, \quad \text{and} \qquad 0 \leq v_j \leq 1, v_0 = 0$$

where $\mu_{E0} - \mu_{C0} + \Delta > 0$ denotes a predefined value from the alternative $\mathrm{H}_{1,\Delta}$ at $stg(0)$, for instance, an a priori guess, and $s_0^2 > 0$ a supposed value for σ^2. An unrealistic small value in (18) may be replaced by some reasonable sample size, for instance, by $n_{\min}$ from (8).

Let us comment the role of the two steering parameters u_j and v_j, $0 \leq u_j \leq 1$ and $0 \leq v_j \leq 1$. By choosing $u_j = 0$ and $v_j = 0$, we get a purely prior information based sample size plan with respect to the parameters. The choice $u_j = 0$ and $v_j > 0$ leads to adaptive plans that only use updated variances, where $s(j)^2$ is the pooled estimator of σ^2 up to $stg(j)$. Such kind of updating is used, for instance, in Denne and Jennison (2000) and references cited therein. For $u_j = 1$, involving $\tilde{n}_i$, the harmonic mean of realized sample sizes, the term $A_j(\Delta)$ is a short-cut version of the meta-analytical combination of standardized mean differences as discussed, for instance, in Hedges and Olkin (1985) and Hartung and Knapp (2001). Putting $u_j = 0$, 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 u_j covers a situation considered in a two-stage-trial by Liu and Chi (2001), and Proschan, Liu, and Hunsberger (2003), who also discuss the role of the standardized mean difference in updating sample sizes.

Let us assume that up to $stg(j-1)$ we have determined sample sizes and weights where $w_\Sigma(j-1) < 1$, by planning with $\Delta_1, \ldots, \Delta_{j-1} \in [0, \Delta_0]$ and at $stg(j)$ we want to plan with Δ_j, that is, we have in mind to reject H_{0,Δ_j}, $\Delta_j \in [0, \Delta_0]$, see (9). With the realized sample sizes n_{E_i} and n_{C_i}, $i = 1, \ldots, j-1$, $j \geq 2$, and defining $Z_0(-\Delta_j) = 0$, we compute the combination statistic up to $stg(j-1)$, see (12),

$$Z_{j-1}(-\Delta_j) = \sum_{i=1}^{j-1} \sqrt{w_i}\, z_i(-\Delta_j)\ , \quad j \geq 1. \tag{19}$$

Supposed we want to obtain a significant result at the next stage by assigning the full remaining weight $1 - w_\Sigma(j-1)$ to this stage. Then, by use of the projected p-value, say $\hat{p}_{j,m}$, the following combination statistic

$$Z_{j,m}(-\Delta_j) = Z_{j-1}(-\Delta_j) + \sqrt{1 - w_\Sigma(j-1)}\, \Phi^{-1}[1 - \hat{p}_{j,m}], \quad j \geq 1, \tag{20}$$

should attain the critical value $\Phi^{-1}(1-\alpha)$, that is,

$$\hat{p}_{j,m} = 1 - \Phi\left[\left(\Phi^{-1}(1-\alpha) - Z_{j-1}(-\Delta_j)\right) / \sqrt{1 - w_\Sigma(j-1)}\right], \qquad j \geq 1. \tag{21}$$

This projected p-value is gained with the (conditional) power $1-\beta$ at $\vartheta + \Delta_j = \sigma A_{j-1}(\Delta_j) > 0$ by choosing the sample size for the next stage j according to (18) as

$$m_j = m_j(\beta) = f_{j-1}(\hat{p}_{j,m}, \beta, \Delta_j), \qquad j \geq 1. \tag{22}$$

In the above procedure, the full weight is used up and stage j is the last one. In case estimates of parameters involved in the trial may not have been stabilized yet, only a part of $m_j(\beta)$ should be used as sample size n_j, that is $n_j = \varepsilon_j \, m_j(\beta)$, with $0 < \varepsilon_j \leq 1$. The remaining weight after stage $(j-1)$ is also divided proportionally to assign the weight $w_j = \varepsilon_j \, (1 - w_\Sigma(j-1))$ at stage j, that is, summarized,

$$w_j = \varepsilon_j \, (1 - w_\Sigma(j-1)), \quad n_j = \varepsilon_j \, m_j(\beta), \quad n_{Ej} = n_{Cj} \approx n_j/2, \quad j \geq 1. \tag{23}$$

The choice of w_j means a proportional partition of the remaining variance of the final $\mathcal{N}(0,1)$-test distribution.

Choosing a smaller power $(1-\beta_j)$, a possible choice of ε_j is provided by

$$\varepsilon_j = \varepsilon_j(\beta_j) = \frac{m_j(\beta_j)}{m_j(\beta)}, \quad m_j(\beta_j) = f_{j-1}(\hat{p}_{j,m}, \beta_j, \Delta_j), \quad \beta \leq \beta_j < 1, \; j \geq 1. \tag{24}$$

Note that β_j is only a lower bound of the type II error rate in stage j as long as $w_j < 1 - w_\Sigma(j-1)$. A similar basic idea is discussed by Hartung (2001) and applied in a 3-stage Self-designing clinical trial with normal outcomes in Hartung (2006).

The pivotal element ε_j of steering the whole Self-designing process may also be defined in a more direct way. From stage $(j-1)$ we have the p-value $p_{j-1} = p_{j-1}(-\Delta_{j-1}) = 1 - F_{t(n_{j-1}-2)}(T_{j-1}(-\Delta_{j-1}))$ based on n_{j-1} observations. Before realizing stage $(j-1)$, upon the information up to stage $(j-2)$, we can compute the significance level α_{j-1}, which our test statistic should reach in stage $(j-1)$ with probability $1-\beta$, that is,

$$\alpha_{j-1} = x \text{ where } x \text{ solves: } n_{j-1} = f_{j-2}(x, \beta, \Delta_{j-1}), \quad j \geq 2. \tag{25}$$

Comparing this expected value with the observed value, we come to new learning rules for n_j and w_j by the following choice of the pivot ε_j as

$$\varepsilon_j = \varepsilon_j^* = r_{Rel} \left(1 - \frac{|\alpha_{j-1} - p_{j-1}|}{\alpha_{j-1} + p_{j-1}} \right) \qquad \text{for } j \geq 2, \tag{26}$$

where r_{Rel} denotes some relaxation factor, $0 < r_{Rel} \leq 1$.

In the extreme cases, when p_{j-1} tends to 1, whereas α_{j-1} is small, or when p_{j-1} tends to 0, the pivot ε_j^* comes near 0. This has the consequence, that n_{min} and w_{min} would be taken for the next stage, see the detailed rules given below. A cautious choice of the relaxation factor is $r_{Rel} = 1/2$, which even in the ideal case, when $\alpha_{j-1} = p_{j-1}$, suggests to take only a half of the remaining weight $1 - w_\Sigma(j-1)$ and sample size $m_j(\beta)$, respectively, for the following stage. For $j = 1$, we may choose ε_1^* as $\varepsilon_1(\beta_1)$ from (24).

Incorporating the minimum sample size and minimum weight introduced in (8), we can formulate the suitably combined learning rules for updating sample sizes and weights as follows: Assume that up to stage $j-1$, $j \geq 1$, there holds

$$n_i \geq n_{\min},\ w_{\min} \leq w_i,\ i=1,\ldots,j-1, \text{ and } w_\Sigma(j-1) = \sum_{i=1}^{j-1} w_i \leq 1 - w_{\min}, \qquad (27)$$

and let ε_j be defined, for instance, by (24) or (26), then, using (22), calculate the weight function

$$W_j = \max\left\{ w_{\min}\ ,\ [1 - w_\Sigma(j-1)]\ \max\left(\varepsilon_j, \frac{n_{\min}}{m_j(\beta)}\right)\right\}, \qquad (28)$$

and set the weight w_j and the sample size n_j of the next stage j as follows:

$$w_j = \begin{cases} W_j & , \quad \text{if } 1 - W_j - w_\Sigma(j-1) \geq w_{\min}, \\ 1 - w_\Sigma(j-1) & , \quad \text{otherwise, and put } j = k, \end{cases} \qquad (29)$$

and

$$n_j = \max\left\{ n_{\min}\ ,\ \frac{w_j}{1 - w_\Sigma(j-1)}\ m_j(\beta)\right\}. \qquad (30)$$

The choice of w_j in (29) and n_j in (30) guarantees the conditions in (27) for all stages and thus, in particular, the upper boundary for the number of performed stages is $1/w_{\min}$. Moreover, the full power $1-\beta$ is reached latest in stage $j = k$, conditioned on $\vartheta + \Delta_k = \sigma A_{k-1}(\Delta_k) > 0$.

5 A confidence interval on the variance parameter

Let $F_{\chi^2(\nu)}$ denote the distribution function of a χ^2-variate with ν degrees of freedom. With the χ^2-statistics from (4), we have in analogy to (11)

$$F_{\chi^2(n_i-2)}\left((n_i-2)\frac{s_i^2}{\sigma^2}\right) \sim \mathcal{U}(0,1), \qquad n_i = n_{Ei} + n_{Ci}, \quad i = 1,\ldots,k, \qquad (31)$$

leading to the combination statistic

$$Z_k^V(\sigma^2) = \sum_{i=1}^{k} \sqrt{w_i}\ \Phi^{-1}\left[F_{\chi^2(n_i-2)}\left((n_i-2)\frac{s_i^2}{\sigma^2}\right)\right] \ \sim\ \mathcal{N}(0,1), \quad \sum_{i=1}^{k} w_i = 1, \qquad (32)$$

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

Often the predefined confidence level for the variance parameter is lower than the one for the outcome measure. So, let us denote the confidence level for the variance by $1-2\kappa$, $0 < \kappa < 1/2$. With the unique solutions of the equations

$$Z_k^V(\sigma_L^2) = \Phi^{-1}(1-\kappa) \qquad \text{and} \qquad Z_k^V(\sigma_U^2) = -\Phi^{-1}(1-\kappa),$$

we build the $(1-2\kappa)$-confidence interval

$$\mathrm{VCI}(\sigma^2) = [\sigma_L^2, \sigma_U^2]. \tag{33}$$

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

6 An example for the effect measure difference of means showing switching from non-inferiority to superiority

Let us consider a clinical examination in which a new agent in an experimental group E is compared to a control group C. The response variables are assumed as (essentially) normally distributed. Let the parameter of interest ϑ be the difference of means, say $\vartheta = \mu_E - \mu_C$, and for both groups a common variance σ^2 is assumed. In such a controlled clinical trial concerning patients with acne papulopustulosa, Lehmacher and Wassmer (1999) discuss an adaptive 3-stage group sequential test of Pocock (1977) type, which led to an early stop for superiority of E with respect to C after the second stage at the one-sided overall significance level of $\alpha = 0.005$. The response variable is the reduction of bacteria (after 6 weeks of treatment) from baseline, examined on agar plates and measured as log CFU / cm^2, CFU: colony forming units. We have taken over the parameter estimates as presented in Table 1. The non-inferiority margin may be predefined as $\Delta_0 = 0.1$.

The test level is also chosen as $\alpha = 0.005$ and the power as $1-\beta = 0.80$. Each stage is planned with equal sample sizes in both groups. Planning with $\Delta_1 = 0.1$ for showing non-inferiority, we get the prior guess $A_0(\Delta_1) = 0.9$ using the prior guesses of ϑ and σ from Table 1. With the critical value $\Phi^{-1}(0.995) = 2.576$, we obtain the total sample size for a one-stage trial using (18),

$$m_1 = f_0(0.005,\ 0.2,\ 0.1) = 57.6.$$

Note that, for the superiority test with $\Delta = 0$, we would calculate the total sample size as 73.

It was intended to start with a $(1/3)m_1$, but by randomizing medications in blocks of size 6, the first sample was chosen to have the size $n_1 = 24$, that is, $\varepsilon_1 = n_1/m_1 = 0.4 = w_1$, see (23). The trial starts and we obtain $y_1 = 1.549$ and $s_1 = 1.316$, leading to the small

Table 1: Self-designing two-stage clinical trial concerning patients with acne papulopustulosa: Data and confidence intervals on the treatment difference $\vartheta = \mu_E - \mu_C$ and on the standard deviation σ.

Stage	Adaptive sample size	Adaptive weight	Treatment difference	Standard deviation	p-value $p_i(-\Delta)$	
0	—	—	0.8	1.0	$p_i(-0.1)$	$p_i(0)$
1	24	$\sqrt{0.4} = 0.63$	1.549	1.316	0.0028	0.0043
2	12	$\sqrt{0.6} = 0.77$	1.580	1.472	0.0381	0.0463
Confidence interval on						
		$\mu_E - \mu_C$			σ	
		$[\,0.231\,,\ 2.894\,]$			$[\,1.157\,, 1.797\,]$	
		Confidence level: $1 - 2\alpha = 0.99$			Confidence level: $1 - 2\kappa = 0.90$	

p-value $p_1(-0.1) = 0.0028$. Consequently, we decide to switch in the planning over to showing superiority. That means, we choose now $\Delta_2 = 0$.

At first we have to compute $Z_1(-\Delta_2) = \sqrt{0.4}\,\Phi^{-1}(1 - 0.0043) = 1.66$ and then the projected p-value, see (21),

$$\hat{p}_{2,m} = 1 - \Phi[(2.576 - 1.66)/\sqrt{0.6}] = 1 - \Phi[1.18],$$

leading to, see (22), with $\Delta_2 = 0$,

$$m_2 = \frac{4\,[1.18 + 0.84]^2}{(1.549\,/\,1.316)^2} = 11.7.$$

We put $u_j = 1$ in (18) because the prior guesses turned out as too cautious. So it was decided to finish the trial by assigning the full remaining weight to the second stage, $w_2 = 0.6$, and to choose the sample size $n_2 = 12$.

By the results of the second stage, see Table 1, we obtain

$$Z_2(0) = 0.63 \cdot 2.63 + 0.77 \cdot 1.68 = 2.95 > 2.576,$$

and equating $Z_2(\vartheta)$ to 2.576 and to -2.576 gives the lower and upper bound, respectively, of the 99%-confidence interval $\mathrm{CI}(\vartheta)$, that is, $\mathrm{CI}(\vartheta) = [0.231, 2.894]$, see also Figure 2 for a graphical display.

For the confidence interval on the variance and the standard deviation, respectively, we choose $\kappa = 0.05$ and obtain $\mathrm{VCI}(\sigma^2)$ by equating

$$Z_2^V(\sigma^2) = \sqrt{0.4}\,\Phi^{-1}\left[F_{\chi^2(22)}\left(22 \cdot \frac{1.316^2}{\sigma^2}\right)\right] + \sqrt{0.6}\,\Phi^{-1}\left[F_{\chi^2(10)}\left(10 \cdot \frac{1.472^2}{\sigma^2}\right)\right]$$

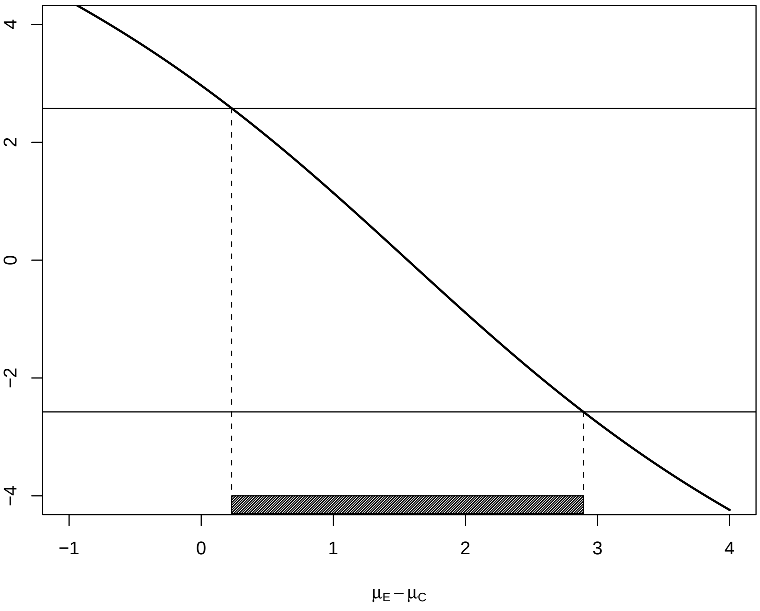

Figure 2: Construction principle of the final 99%-confidence interval for the difference of means $\mu_E - \mu_C$ in the real-data example from Section 6.

to ±1.645. The solutions are $\mathrm{VCI}(\sigma^2) = [1.339, 3.228]$ so that the resulting 90%-confidence interval on σ is given as $\mathrm{VCI}^{1/2}(\sigma) = [1.157, 1.797]$.

7 A confidence interval for the ratio of means and adaptive planning

Let us assume that the independent random variables x_E and x_C, introduced in Section 2, have positive means, $\mu_E > 0$ and $\mu_C > 0$. The same should hold for the observed means, $\bar{x}_{Ei} > 0$, $\bar{x}_{Ci} > 0$, $i = 1, \ldots, k$. The parameter of interest considered now is the ratio of means,

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

Let $\Delta_0 \geq 0$ be again a non-inferiority margin, we test

$$\mathrm{H}^r_{0,\Delta} : \lambda \leq 1 - \Delta \quad \text{versus} \quad \mathrm{H}^r_{1,\Delta} : \lambda > 1 - \Delta, \quad 0 \leq \Delta \leq \Delta_0, \quad \Delta_0 < 1, \tag{34}$$

at a given level α, $0 < \alpha < 1/2$, where $\mathrm{H}^r_{1,\Delta}$ means superiority when $\Delta = 0$, otherwise $(\Delta-)$non-inferiority of E with regard to C.

Following an idea of Fieller (1940), see also Finney (1964), let us introduce the statistics

$$\bar{x}_i(\lambda) = \bar{x}_{Ei} - \lambda\,\bar{x}_{Ci} \;\sim\; \mathcal{N}\left(0\,,\,\left(\frac{1}{n_{Ei}} + \frac{\lambda^2}{n_{Ci}}\right)\sigma^2\right),\; i = 1,\ldots,k, \tag{35}$$

and the t-statistics for $i = 1,\ldots,k$,

$$T_i^r(\lambda) = \frac{\bar{x}_i(\lambda)}{\hat{\sigma}(\bar{x}_i(\lambda))} = \frac{\bar{x}_{Ei} - \lambda\,\bar{x}_{Ci}}{\sqrt{(1/n_{Ei} + \lambda^2/n_{Ci})\,s_i^2}} \;\sim\; t(n_i - 2), \quad n_i = n_{Ei} + n_{Ci}, \tag{36}$$

where s_i^2 is the pooled variance estimator from (3).

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

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}T^r(\lambda) &= \frac{-\,\bar{x}_C\,Q - (\bar{x}_E - \lambda\,\bar{x}_C)\,Q^{-1}\,s^2\,\lambda/n_C}{Q^2}\\ &= \frac{-\,\bar{x}_C\,Q^2 - (\bar{x}_E - \lambda\,\bar{x}_C)\,s^2\,\lambda/n_C}{Q^3}\\ &= \frac{-\,(\bar{x}_C/n_E + \lambda\,\bar{x}_E/n_C)\,s^2}{Q^3} < 0, \qquad \text{for } \lambda > 0.\end{aligned}$$

Hence $T^r(\lambda)$ is monotone decreasing for positive λ. So, we obtain the final weighted inverse normal combination statistic

$$Z_k^r(\lambda) = \sum_{i=1}^{k} \sqrt{w_i}\,\Phi^{-1}\left[F_{t(n_i-2)}(T_i^r(\lambda))\right] \;\sim\; \mathcal{N}(0,1), \quad w_\Sigma(k) = 1\,, \tag{37}$$

which is monotone decreasing in $\lambda, \lambda > 0$.

Defining

$$T_i^r(\infty) = \frac{-\bar{x}_{Ci}}{\sqrt{s_i^2/n_{Ci}}} = \lim_{\lambda\to\infty} T_i^r(\lambda) \quad \text{and} \quad T_i^r(0) = \frac{\bar{x}_{Ei}}{\sqrt{s_i^2/n_{Ei}}},\; i = 1,\ldots,k, \tag{38}$$

and herewith $Z_k^r(\infty)$, $Z_k^r(0)$, we have the following boundaries for $Z_k^r(\lambda)$,

$$Z_k^r(\infty) = \inf_{\lambda>0} Z_k^r(\lambda) < Z_k^r(\lambda) < \sup_{\lambda>0} Z_k^r(\lambda) = Z_j^r(0),\; 0 < \lambda < \infty. \tag{39}$$

In analogy to (14), we can formulate the confidence interval on λ as follows,

$$\mathrm{CI}^r(\lambda) = [\lambda_L\,,\,\lambda_U]\,, \tag{40}$$

where λ_L solves $Z_k^r(\lambda_L) = \Phi^{-1}(1-\alpha)$ if $Z_k^r(0) > \Phi^{-1}(1-\alpha)$, otherwise set $\lambda_L = 0$, and λ_U solves $Z_k^r(\lambda_U) = -\Phi^{-1}(1-\alpha)$ if $Z_k^r(\infty) < -\Phi^{-1}(1-\alpha)$, otherwise set $\lambda_U = \infty$. The unique solutions of (40) can again easily be found iteratively by standard statistics software packages. The confidence coefficient of $\mathrm{CI}^r(\lambda)$ is $1 - 2\alpha$.

For $w_1 = 1 = k$, solving the equations implied by (40) explicitly, we get a formal representation of Fieller's well-known confidence interval for the ratio of means, see Fieller (1940), Finney (1964).

In the test problem (34), we proceed as follows at level α, $0 < \alpha < 1/2$:

$$\begin{aligned} &\text{if } 1-\Delta < \lambda_L, && \text{then reject } \mathrm{H}^r_{0,\Delta}, \\ &\text{if } 1-\Delta_0 \geq \lambda_L, && \text{then stay with } \mathrm{H}^r_{0,\Delta_0}. \end{aligned} \tag{41}$$

In the following, we present some considerations on learning rules for adaptively chosen samples sizes and weights in the present context. Planning with equal sample sizes in the two groups and suppressing the subscript i, we set $n_E = n_C = M$, $\xi = \mu_E - (1-\Delta)\mu_C$ for a fixed $\Delta \in [0\,,\,\Delta_0]$, and $x = x_E - (1-\Delta)x_C$. Then

$$x \sim \mathcal{N}\left(\xi, \sigma(x)^2\right) \quad \text{and} \quad \bar{x} \sim \mathcal{N}\left(\xi, \frac{1}{M}\,\sigma(x)^2\right), \tag{42}$$

where $\sigma(x)^2 = (1+(1-\Delta)^2)\sigma^2$.

For given type I and II error rates α and β, respectively, testing the point hypotheses

$$\mathrm{H}_0^* : \xi = 0 \quad \text{versus} \quad \mathrm{H}_1^* : \xi = \xi^* > 0$$

by

$$T_0^r(1-\Delta) = \sqrt{M}\,\frac{\bar{x}}{\sigma(x)} \;\sim\; \mathcal{N}(0,1) \quad \text{under } \mathrm{H}_0^*, \tag{43}$$

the required sample size M has to be chosen (one-sample formula) as follows,

$$M = \frac{\left[\max\{0\,,\,\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\}\right]^2}{\left(\xi^*/\sigma(x)\right)^2}\,. \tag{44}$$

At $stg(0)$, let $s_0^2 > 0$ be an assumed value for σ^2 and $\xi^* = \mu_{E0} - (1-\Delta)\mu_{C0} > 0$ be a chosen value in the alternative $\mathrm{H}^r_{1,\Delta}$, then the sample size $n = 2M$ for both groups is obtained by the sample size spending function $g_0(\alpha,\beta,\Delta)$ defined by

$$g_0(\alpha,\beta,\Delta) = 2\,\frac{\left[\max\{0\,,\,\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\}\right]^2}{B_0(\Delta)^2}, \tag{45}$$

where

$$B_0(\Delta) = \frac{\mu_{E0} - (1-\Delta)\,\mu_{C0}}{s_0\,\sqrt{1+(1-\Delta)^2}} > 0.$$

Instead of the normal test statistic $T_0^r(\lambda)$ from (43), the t-statistic $T_i^r(\lambda)$ from (36) is used at the i-th stage of the trial and so $g_0(\alpha,\beta,\Delta)$ delivers approximate, lower values for the desired sample sizes. For ease of presentation, we further consider only a purely sample based updating, say by $g_j(\cdot, \beta, \Delta)$, $j \geq 1$, and a mixture between g_0 for exclusively prior information based sample size planning, and g_j, $j \geq 1$, can be arranged in the same kind as demonstrated in Section 4, see (18). We estimate the standardized mean difference, under the alternative $\mathrm{H}^r_{1,\Delta}$, in the denominator of (44) at stage j by combining the estimates of $stg(1)$ up to $stg(j)$ weighted by the harmonic means of the realized sample sizes in the stages. We obtain

$$g_j(\alpha, \beta, \Delta) = 2\, \frac{\left[\max\{0\,,\ \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}\right]^2}{B_j(\Delta)^2}, \quad j = 1, \ldots, k, \tag{46}$$

where

$$B_j(\Delta) = \sum_{i=1}^{j} \frac{\tilde{n}_i}{\sum_{h=1}^{j} \tilde{n}_h}\, \frac{\bar{x}_{Ei} - (1-\Delta)\bar{x}_{Ci}}{s_i\sqrt{1+(1-\Delta)^2}} > 0\,, \quad \tilde{n}_i = \frac{2}{1/n_{Ei} + 1/n_{Ci}}.$$

If B_j is not positive, g_j may be replaced by a part of g_0, or the trial is not continued with the specified non-inferiority margin Δ in mind. An unrealistic small value in (45) or (46) may be replaced, for instance, by $n_{\min}$ from (8).

The test statistic for (34) is $Z^r_k(1-\Delta)$, see (37). Assume that up to $stg(j-1)$ we have gained $Z^r_i(\lambda) = \sum_{h=1}^{i} \sqrt{w_h}\, z^r_h(\lambda)$, with $z^r_h(\lambda) = \Phi^{-1}\left[F_{t(n_h-2)}\left(T^r_h(\lambda)\right)\right]$. Then, in analogy to (21), we derive the projected p-value for $stg(j)$ as

$$\hat{p}^r_{j,m} = 1 - \Phi\left[\left(\Phi^{-1}(1-\alpha) - Z^r_{j-1}(1-\Delta)\right) / \sqrt{(1 - w_\Sigma(j-1))}\right]\,,\ j \geq 1, \tag{47}$$

which as in (22), (23) yields the needed sample size and weight for $stg(j)$ as, see (45), (46),

$$n_j = \varepsilon_j\, m^r_j(\beta) \quad \text{and} \quad w_j = \varepsilon_j\,(1 - w_\Sigma(j-1)),\ 0 < \varepsilon_j \leq 1, \tag{48}$$

where $m^r_j = m^r_j(\beta) = g_{j-1}(\hat{p}^r_{j,m}, \beta, \Delta)$, $n_{Ej} = n_{Cj} \approx n_j/2$, $j = 1, \ldots, k$. The power is conditioned on $\mu_E - (1-\Delta)\mu_C = B_{j-1}(\Delta)\sigma\sqrt{1+(1-\Delta)^2} > 0$. The pivotal learning element ε_j can be chosen in an analogue manner as in (24) and (26). Taking into account a minimum weight and sample size at each stage, see (8), the suitably combined learning rules of (29) for updating sample sizes and weights can be carried over.

8 An example for the ratio of means showing switching from superiority to non-inferiority

Let us consider a clinical trial, one of the authors was concerned with as a biometrical advisor. A new (E) and a standard drug (C), two different inhalers, for treating patients

Table 2: Self-designing clinical trial treating patients with asthma bronchiale for the effect measure ratio of means λ concerning a lungs functioning parameter (FEV$_1$): Data, confidence interval on λ with confidence coefficient 0.95, and combined test statistics.

	Sample		Data (in ℓ)			Confidence interval	Combined	
Stage	size	Weight	on			on	test	
i	n_i	$\sqrt{w_i}$	μ_E	μ_C	σ	$\lambda = \mu_E/\mu_C$	statistics	
0	—	—	2.75	2.50	0.75	$1 - \Delta_0 = 0.90$	$Z_i^r(1.0)$	$Z_i^r(0.9)$
1	128	$\sqrt{1/3}$	2.67	2.55	0.81		0.482	1.563
2	56	$\sqrt{2/3}$	2.70	2.56	0.87	[0.951 , 1.162]	0.971	2.997

with asthma bronchiale are compared with respect to a lung function parameter named FEV$_1$: forced expiratory volume in 1 second, measured in liter (ℓ). The ratio of means is the common outcome measure in that application. A nearly normed non-inferiority margin for the clinical parameter is $\Delta_0 = 10\%$. The type I and II error rates of the trial are chosen as $\alpha = 0.025$ and $\beta = 0.10$, respectively. The two treatment groups are equally sized at each stage and the drugs are equally randomized within blocks of size 8. The investigators were optimistic so that the trial starts with an attempt to show superiority ($\Delta = 0$). The first weight is scheduled as $w_1 = 1/3$ or $\varepsilon_1 = 1/3$.

The critical value is 1.96 and, with the assumed prior information from Table 2, we compute by (45) for a one-stage trial 378 patients to be observed (with $\Delta = 0$). Using (48) we obtain $n_1 = 126$ and choose $n_1 = 128$ because of the randomization scheme. With the observed data, see Table 2, we obtain $Z_1^r(1) = 0.48$. We recognize that the prior guesses of the parameters were too optimistic for that study population with respect to the new drug. So we use in the planning for the next stage only the observed values of $stg(1)$, especially as they are based on a relatively large number of patients. So, with $\Delta = 0$, we get $\hat{p}_{2,m}^r = 0.035$, $B_1(0) = 0.105$, and, for m_2^r with $\Delta = 0$, we calculate a number of 1736 patients to be observed at the next stages for the chance of showing superiority. So the decision was made to stay with showing non-inferiority being sufficient for regulatory concerns.

Planning with $\Delta = \Delta_0 = 0.10$, we compute $Z_1^r(0.90) = 1.56$, $\hat{p}_{2,m}^r = 0.31$, $B_1(0.10) = 0.344$ and by (46) for the total size of the remaining stages $m_2^r = 53$. It was decided to finish the trial after the second stage, so the final sample size is $n_2 = 56$ because of the randomization scheme. The combination statistic is $Z_2^r(\lambda) = \sqrt{1/3}\; z_1^r(\lambda) + \sqrt{2/3}\; z_2^r(\lambda)$, see (37). Equating $Z_2^r(\lambda)$ to ± 1.96 and solving for λ leads to the confidence interval,

$\mathrm{CI}^r(\lambda) = [0.951\ ,\ 1.162]$, on the ratio λ, which lies clearly above 0.90. Further, with $z_2^r(0.90) = 1.76$, we calculate the final test value, $Z_2^r(0.90) = 0.58 \cdot 2.70 + 0.82 \cdot 1.76 = 3.01 > 1.96$, confirming significant non-inferiority.

9 Final remarks

Confidence intervals on the effect measures difference and ratio of means are derived by combining parameterized t-statistics via the weighted inverse normal method. Assigning consecutively different weights to the stages, the number of stages is determined during the ongoing trial. Suitably combined learning rules are derived for simultaneously updating sample sizes and weights. The consequence is an effective controlling of the clinical trial, see also Fisher (1998) for general considerations in that direction.

The impression may arise that Self-designing concepts are a matter more for longer running studies with many interim analyses. But let us consider a situation where, based on the available a priori information, a two-stage trial seems to be appropriate. Usually no surprising positive results are expected in the interim analysis, so that in the most practical applications, an O'Brien and Fleming (1979) design is chosen, that provides a greater chance for showing significance at the end of the study than, for instance, the Pocock (1977) design. However, there is practically no chance to show significance in the interim analysis. For example, an one-sided O'Brien and Fleming test at overall level $\alpha = 0.025$ needs for significance a level attained at the end of the study of 0.024, but of 0.0026 in the interim analysis. So in that situation, a better choice would be a Self-designing concept, where the weight for the first stage can be set to 1/2 as in the usual 2-stage O'Brien and Fleming design. Then the full level α is preserved at the end of the study, but we have the additional option to decide in the interim analysis for at least one further interim analysis if the observed treatment effects will not satisfy the expectations.

Choosing in advance a 3-stage O'Brien and Fleming design, is not a good idea in the considered situation, because then, even in the second stage, a low level attained of only 0.007 would be needed for showing significance. So nearly surely, a third stage could not be avoided. For comparison, the corresponding Pocock design needs a level attained of 0.011 at each of the three planned stages, whereas the Self-designing concept needs just the full level of 0.025 to be attained at the end of the study after one or more interim analyses.

Consequently, a Self-designing concept can be a reasonable alternative to classical group sequential trials, see also the simulation results reported in Hartung (2006), and the real-data examples in Section 6 and 8. Moreover, Self-designing can be considered as the limit case of O'Brien and Fleming designing, when the needed level attained assigned to the last stage of the trial tends to the full overall significance level, as discussed by Hartung (2006). That corresponds, in the Wang and Tsiatis (1987) δ-class of group sequential trials, to the limit case when the design parameter δ tends to $-\infty$. In a non-adaptive setting, this makes less sense. But in an adaptive approach, interim analyses are used not only for considering safety concerns of the clinical trial but also for the chance to reassess the sample size planning, and being not less important, the number of possible interim analyses has not to be specified in advance anymore.

Besides all these considerations, in spite of its vital practical importance, the effect measure ratio of means, with the variances of the outcomes assumed to be known or not, seems not to be considered as well in classical group sequential trials as in their adaptive extensions until now, neither for testing non-inferiority nor for deriving confidence intervals.

Sample sizes n are computed in Sections 4 and 7 through a normal approximation for applying a $t(n-2)$-variate. Nearly exact values are achieved by correcting with the variance of a $t(n-2)$-variate, that is, replacing n by $n_{\text{corr}} = n(n-2)/(n-4)$, $n \geq 5$, being relevant for small values of n. The idea behind is the same as in replacing a t-variate by a normal variate with identical variance. However, computed values usually have to be modified to take into account the particular randomization scheme applied in a clinical trial.

Unlike the inverse chi-square (χ^2) combination method considered, for instance, by Bauer and Köhne (1994), Liu and Chi (2001), and Frick (2002) in two-stage designs and by Hartung (2000) and Hartung and Knapp (2003, 2006) for Self-designing trials, the inverse normal combination method is symmetric in the sense that positive values of the t-statistics are accumulated in the same way as negative values. So no direction of deviations from the null-distribution is preferred, see also Hedges and Olkin (1985, p. 40). Even when sample sizes and weights of the stages are identical, the results by applying both combination methods to the same data may differ. For instance, in the real-data example of Self-designing discussed in Hartung (2006, p. 523), combining by use of the inverse normal method yields a global p-value (0.0027) that is less than a half of the global p-value (0.0057) reached by applying the inverse χ^2 method to the same observed data of the three stages when testing for superiority. This tendency is in concordance with simulation results which assign a higher mean sample size to the inverse χ^2-method in order to reach the same p-values as the inverse normal method, see Hartung (2006).

Finally let us briefly address point estimation. The combination statistic $Z_k(\vartheta)$ from Section 3 is $\mathcal{N}(0,1)$-distributed with mode and median 0. A maximum likelihood (ML) estimator $\hat{\vartheta}_{ML}$ of the difference $\vartheta = \mu_E - \mu_C$ is given as the solution of $Z_k(\hat{\vartheta}_{ML}) = 0$. Sometimes, such an estimator is also called pseudo ML-estimator. The *global* p-value is $p_G(\vartheta) = 1 - \Phi(Z_k(\vartheta))$, and solving the equation $p_G(\vartheta) = 1/2$ yields $\hat{\vartheta}_{ML}$ as solution. Hence, noting that $Z_k(\vartheta)$ is monotone in ϑ, $\hat{\vartheta}_{ML}$ is median unbiased, cf. Cox and Hinkley (1974, p. 273), Liu and Chi (2001). That means, the ML-estimator lies with equal probability below and above the parameter ϑ. For large sample sizes n_i, $\hat{\vartheta}_{ML}$ is approximated by

$$\hat{\vartheta}^A_{ML} = \sum_{i=1}^{k} \Big[y_i \sqrt{w_i}/\hat{\sigma}(y_i) \Big] \Big/ \Big[\sum_{h=1}^{k} \sqrt{w_h}/\hat{\sigma}(y_h) \Big],$$

see (16) and (17), which uses the inverse estimated standard errors instead of the inverse estimated variances of the y_i's as known from meta-analysis, see Hartung, Knapp, and Sinha (2008). Weighted means like $\hat{\vartheta}^A_{ML}$ are used in the generalized Cochran-Wald statistics considered by Hartung, Böckenhoff, and Knapp (2003).

Using $Z_k^V(\sigma^2)$ from Section 5 yields the median unbiased ML-estimator $\hat{\sigma}^2_{ML}$ of σ^2 by solving $Z_k^V(\hat{\sigma}^2_{ML}) = 0$, and via $Z_k^r(\lambda)$ from Section 7, we get the median unbiased ML-estimator $\hat{\lambda}_{ML}$ of the ratio $\lambda = \mu_E/\mu_C$ as the solution of $Z_k^r(\hat{\lambda}_{ML}) = 0$.

References

[1] Bauer, P. and Kieser, M. (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika* **83,** 934–937.

[2] Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50,** 1029–1041.

[3] Brannath, W., Bauer, P., Maurer, W., and Posch, M. (2003). Sequential tests for noninferiority and superiority. *Biometrics* **59,** 106–114.

[4] Brannath, W., König, F., and Bauer, P. (2003). Improved repeated confidence bounds in trials with a maximal goal. *Biometrical Journal* **45,** 311–324.

[5] Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97,** 236–244.

[6] Cheng, Y. and Shen, Y. (2004). Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics* **60,** 910–918.

[7] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* New York: Chapman and Hall.

[8] Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine* **20,** 2645–2660.

[9] Denne, J. S. and Jennison, C. (2000). A group sequential t-test with updating of sample sizes. *Biometrika* **87,** 125–134.

[10] EMEA (The European Agency for the Evaluation of Medicinal Products) (2000). *Points to Consider on Switching between Superiority and Non-inferiority.* London, CPMP/EWP/482/99.

[11] Fieller, E. C. (1940). The biological standardization of insulin. *Journal of the Royal Statistical Society* (Suppl.) **7,** 1–64.

[12] Finney, D. J. (1964). *Statistical Methods in Biological Assay.* 2$^{\text{nd}}$ Edition. London: Griffin.

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

[14] Frick, H. (2002). On confidence bounds for the Bauer-Köhne two-stage test. *Biometrical Journal* **44,** 241–249.

[15] Hartung, J. (2000). A new class of self-designing clinical trials. In: Hasman A. *et al.* (eds.) *Medical Infobahn for Europe.* Proceedings of MIE 2000 and GMDS 2000. IOS Press, Amsterdam, 310–314.

[16] Hartung, J. (2001). A self-designing rule for clinical trials with arbitrary response variables. *Controlled Clinical Trials* **22,** 111–116.

[17] 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.

[18] Hartung, J., Böckenhoff, A., and Knapp, G. (2003). Generalized Cochran-Wald statistics in combining of experiments. *Journal of Statistical Planning and Inference* **113,** 215–237.

[19] Hartung, J. and Knapp, G. (2001). On tests of the overall treatment effect in the meta-analysis with normally distributed responses. *Statistics in Medicine* **20,** 1771–1782.

[20] Hartung, J. and Knapp, G. (2003). A new class of completely self-designing clinical trials. *Biometrical Journal* **45,** 3–19.

[21] Hartung, J. and Knapp, G. (2006). Repeated confidence intervals in self-designing clinical trials and switching between noninferiority and superiority. *Biometrical Journal* **48,** 697–709.

[22] Hartung, J., Knapp, G., and Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications*. New York: Wiley.

[23] Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Orlando: Academic Press.

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

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

[26] Jennison, C. and Turnbull, B. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: CRC Press Inc.

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

[28] Liu, Q. and Chi, G. Y. H. (2001). On sample size and inference for two–stage adaptive designs. *Biometrics*, **57,** 172–177.

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

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

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

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

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

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