Adaptive Controlled Noninferiority Group Sequential Trials

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

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

Abstract: For studies comparing three independent arms: test group T, reference group R, and control group C, we consider the hierarchical testing of the a priori ordered hypotheses, that, in short,

$$\begin{aligned} \text{(I)}: &\quad T > C, \\ \text{(II)}: &\quad T > R - \Delta, \quad \Delta > 0, \end{aligned}$$

in general adaptive group sequential designs. For normally distributed response variables with unknown variances, nested confidence intervals on the study parameters are derived at each stage of the trial, holding a predefined confidence level. During the course of the trial, the sample sizes can be calculated in a completely adaptive way based on the unblinded data of previous stages. Concrete formulae for sample size updating are provided in this paper. Moreover, in each interim analysis, it is possible to switch in the planning from showing noninferiority of T in (II) to showing superiority of T, that is, $T > R$.

A real data example is worked out in detail following an adaptive three-stage design of Pocock (1977) type. In the example, (I) is shown in the first stage and (II) in the second stage, so that the study stopped earlier at the second stage.

Keywords: Controlled noninferiority trials; Hierarchical testing; Group sequential confidence intervals; Adaptive sample size planning; Switching from noninferiority to superiority

1 Introduction

Several clinical trial guidelines, see for instance EMEA (1998), recommend to include a placebo control group C, when an experimental test group T is to be compared to a standard reference group R with respect to noninferiority. A more detailed regulatory point of view is formulated by Koch (2006), who essentially says, that in areas, where

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

difficulties exist with the description of the patient population in such a way that placebo-response and response under a standard treatment can be well predicted, it may be necessary to include both the placebo and active comparator in the confirmatory phase III trial. It is an ethical mandate that the number of patients randomized to the placebo comparison be limited as much as possible. An adaptive design combined with a multiple testing procedure may offer the opportunity to stop recruitment to the placebo group after an interim analysis, as soon as superiority of the experimental treatment over placebo has been demonstrated. The trial is then continued into further stages to demonstrate the noninferiority of the experimental treatment in comparison to the reference treatment.

By these considerations, we have a good description of the subject of the present paper. With some noninferiority margin $\Delta > 0$, we test the a priori ordered hypotheses, that, in short,

$$\begin{aligned} \text{(I)}: &\quad T > C, \\ \text{(II)}: &\quad T > R - \Delta, \quad \Delta > 0. \end{aligned}$$

When (I) is shown, we can test for (II). This proceeding has the positive consequence, that for both hypotheses tests, we can take the same significance level, that describes the overall test level, too. A controlled noninferiority trial is considered, for instance, also by Pigeot et al. (2003), who present, in a one-stage trial, a different approach, where in a first step it has to be shown that: $R > C$. Only when being here successful, other comparisons are allowed. That approach bears the risk, that the whole study breaks down, when R fails to be superior to C.

Excluding that risk, we may add in our approach at third order (III): $R > C$. But the interest of the study is directed towards T, so that R is less important, especially as R is usually well established on the market, which, however, does not imply to be very effective.

In this paper, we consider normally distributed response variables, with unknown variances, in general adaptive group sequential trials, see Hartung (2006). Parameterized p-values, see Cox and Hinkley (1974), of the several stages are combined by use of the inverse normal or Stouffer's method, well known from meta-analysis, see Hartung, Knapp, and Sinha (2008, Chapter 3). The resulting combined statistics are used for group sequential hierarchical testing of the a priori ordered hypotheses (I) and (II). A test on the homogeneity of the stage specific treatment effects is derived. Further, the concept of repeated confidence intervals, see Jennison and Turnbull (2000) and references cited therein,

is extended, in an exact way, to the case of unknown variances and possibly adaptively chosen sample sizes. Moreover, in the considered adaptive sequential situation, where the end of the study depends on a test decision, median unbiased maximum likelihood estimators of the study parameters can be derived, including the possibly different variance parameters.

In each interim analysis, it is possible to change the planning from showing noninferiority of T to showing superiority of T with regard to R, too. We present a group sequential confidence interval approach to switching from noninferiority to superiority, see Bauer and Kieser (1996) and, for instance, the clinical trial guideline EMEA (2000).

Further, we develop formulae for sample size calculation in group sequential trials. These formulae seem to be unknown so far, even in case of non-adaptive group sequential trials, where the computed sample size for the first stage is taken in all following stages.

The outline of the present paper is as follows: In Section 2, the hierarchical testing of the a priori ordered hypotheses is developed and the homogeneity of the stage specific treatment differences is tested. Section 3 contains group sequential confidence intervals on the treatment differences and the model parameters. Section 4 presents median unbiased maximum likelihood estimators and meta-analytical estimators of the treatment effects and of the model parameters. Section 5 contains the formulae for sample size calculation and rules for adaptively updating the sample sizes. Section 6 presents a real data example, following an adaptive three-stage design of Pocock (1977) type, in detail. There also, the added test of (III): $R > C$ is discussed in connection with the example. Section 7 contains some further comments, especially concerning the choice of the critical values.

2 Group Sequential Testing

Let us consider a new treatment in a test group T, a standard treatment in a reference group R, and a placebo treatment in a control group C. The associated response variables may be denoted by X_T, X_R, and X_C, which are mutually stochastically independent normally distributed random variables with means μ_T, μ_R, μ_C and variances $\sigma_T^2 > 0$, $\sigma_R^2 > 0$, and $\sigma_C^2 > 0$, respectively, that is,

$$X_T \sim \mathcal{N}\left(\mu_T, \sigma_T^2\right), \quad X_R \sim \mathcal{N}\left(\mu_R, \sigma_R^2\right), \quad X_C \sim \mathcal{N}\left(\mu_C, \sigma_C^2\right).$$

At level α, $0 < \alpha < 1/2$, we first test whether T is superior to C, that ist, we test the hypotheses:

$$H_0^{TC} : \mu_T = \mu_C \quad \text{versus} \quad H_1^{TC} : \mu_T > \mu_C. \tag{1}$$

If H_o^{TC} is rejected at level α in favour of H_1^{TC}, then we test, at the same level α, the noninferiority hypotheses of T with regard to R,

$$H_0^{TR} : \mu_T = \mu_R - \Delta \quad \text{versus} \quad H_1^{TR} : \mu_T > \mu_C - \Delta,\ \Delta \in [0, \Delta_0], \tag{2}$$

where $\Delta_0 \geq 0$ denotes some margin for the noninferiority parameter Δ. This hierarchical testing procedure holds the overall significance level α, see Maurer, Hothorn, and Lehmacher (1995), and, for instance, Pigeot et al. (2003) for an application of this test principle.

We consider a comparative study, which is carried out in a number of independent stages, say K. In the i-th stage, $i = 1, \ldots, K$, let be $\bar{X}_{T_i}$, $\bar{X}_{R_i}$, and $\bar{X}_{C_i}$ the sample means of $n_{T_i} \geq 2$, $n_{R_i} \geq 2$, and $n_{C_i} \geq 2$ responses in the respective treatment groups. The variance parameters can be estimated by the corresponding sample variances $S^2_{T_i}$, $S^2_{R_i}$, and $S^2_{C_i}$, which are stochastically independent of the means and follow scaled χ^2-distributions, that is, for $i = 1, \ldots, K$,

$$(n_{T_i} - 1)\,\frac{S^2_{T_i}}{\sigma^2_T} \sim \chi^2_{n_{T_i}-1}, \quad (n_{R_i} - 1)\,\frac{S^2_{R_i}}{\sigma^2_R} \sim \chi^2_{n_{R_i}-1}, \quad (n_{C_i} - 1)\,\frac{S^2_{C_i}}{\sigma^2_C} \sim \chi^2_{n_{C_i}-1}. \tag{3}$$

2.1 Test Statistics

The parameters of interest are $\theta_{TC} = \mu_T - \mu_C$ and $\theta_{TR} = \mu_T - \mu_R$. Denote t_ν the central t-distribution with ν degrees of freedom, then with the true parameters θ_{TC} and θ_{TR}, we have, in good approximation, at the i-th stage, $i = 1, \ldots, K$,

$$D_i^{TC}(\theta_{TC}) := \frac{\bar{X}_{T_i} - \bar{X}_{C_i} - \theta_{TC}}{\sqrt{\hat{\sigma}^2_{T_i}/n_{T_i} + \hat{\sigma}^2_{C_i}/n_{C_i}}} \sim t_{\nu_i(TC)}, \tag{4}$$

$$D_i^{TR}(\theta_{TR}) := \frac{\bar{X}_{T_i} - \bar{X}_{R_i} - \theta_{TR}}{\sqrt{\hat{\sigma}^2_{T_i}/n_{T_i} + \hat{\sigma}^2_{R_i}/n_{R_i}}} \sim t_{\nu_i(TR)}, \tag{5}$$

where

$$\hat{\sigma}^2_{T_i} = S^2_{T_i}, \quad \hat{\sigma}^2_{C_i} = S^2_{C_i}, \quad \hat{\sigma}^2_{R_i} = S^2_{R_i},$$

and with Satterthwaite's approximation,

$$\nu_i(TC) = \frac{\left(S^2_{T_i}/n_{T_i} + S^2_{C_i}/n_{C_i}\right)^2}{\left(S^2_{T_i}/n_{T_i}\right)^2/(n_{T_i} - 1) + \left(S^2_{C_i}/n_{C_i}\right)^2/(n_{C_i} - 1)},$$

$$\nu_i(TR) = \frac{\left(S_{T_i}^2/n_{T_i} + S_{R_i}^2/n_{R_i}\right)^2}{\left(S_{T_i}^2/n_{T_i}\right)^2/(n_{T_i}-1) + \left(S_{R_i}^2/n_{R_i}\right)^2/(n_{R_i}-1)}.$$

Provided $\sigma_T^2 = \sigma_C^2$, then both parameters are estimated in the i-th stage by the pooled estimator

$$\hat{\sigma}_{T_i}^2 = \hat{\sigma}_{C_i}^2 = \frac{(n_{T_i}-1)S_{T_i}^2 + (n_{C_i}-1)S_{C_i}^2}{n_{T_i} - n_{C_i} - 2}, \tag{6}$$

and in (4), we get an exact t-distribution with $\nu_i(TC) = n_{T_i} + n_{C_i} - 2$ degrees of freedom, $i = 1, \ldots, K$. Analogously, we proceed when $\sigma_T^2 = \sigma_R^2$.

If $\sigma_T^2 = \sigma_C^2 = \sigma_R^2 =: \sigma^2$, then the common variance is estimated in the i-th stage by

$$\hat{\sigma}_i^2 = \frac{(n_{T_i}-1)S_{T_i}^2 + (n_{C_i}-1)S_{C_i}^2 + (n_{R_i}-1)S_{R_i}^2}{n_{T_i} + n_{C_i} + n_{R_i} - 3} \tag{7}$$

and in (4) and (5), $\nu_i(TC) = \nu_i(TR) = n_{T_i} + n_{C_i} + n_{R_i} - 3$, $i = 1, \ldots, K$.

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

$$1 - p_i^d(\theta_d) = F_{t_{\nu_i(d)}}(D_i^d(\theta_d)) \sim \mathcal{U}(0,1), \quad d = TC, TR,\ i = 1, \ldots, K, \tag{8}$$

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

$$z_i^d(\theta_d) := \Phi^{-1}(1 - p_i^d(\theta_d)) \sim \mathcal{N}(0,1), \quad d = TC, TR,\ i = 1, \ldots, K, \tag{9}$$

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

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

$$Z_j^d(\theta_d) := \sum_{i=1}^{j} z_i^d(\theta_d) \sim \sqrt{j}\ \mathcal{N}(0,1), \quad d = TC, TR,\ j = 1, \ldots, K. \tag{10}$$

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

$$P\left(\sum_{i=1}^{j} Y_i \le cv_j(d) \text{ for all } j = 1, \ldots, K\right) = 1 - \alpha, \quad d = TC, TR, \tag{11}$$

see Hartung (2006), and a respective comment in Section 7.

Using these critical values $cv_j(d)$, we get the following probability statements for the combining pivotal statistics from (10),

$$P_{\theta_d}\Big(Z_j^d(\theta_d) \le cv_j(d) \text{ for } j = 1, \ldots, k \le K\Big) \begin{cases} \ge 1-\alpha \text{ for } k < K, \\ = 1-\alpha \text{ for } k = K, \end{cases} \qquad (12)$$

$$d = TC, TR.$$

Consequently, we can formulate the following test procedure at overall level of at most α as implied by (12): At the k-th stage, $k = 1, \ldots, K$, we reject H_0^{TC} in favour of H_1^{TC}, see (1),

$$\text{if } \exists i_0 \in \{1, \ldots, k\} : Z_{i_0}^{TC}(0) > cv_i(TC). \qquad (13)$$

Provided the decision is made for the alternative H_1^{TC}, then in the noninferiority test problem (2), we decide in favour of the alternative $H_{1,\Delta}^{TR}$, $\Delta \in [0, \Delta_0]$,

$$\text{if } \exists j_\Delta \in \{1, \ldots, k\} : Z_{j_\Delta}^{TR}(-\Delta) > cv_j(TR). \qquad (14)$$

Provided the placebo arm C is not stopped after stage i_0, since each stage of the trial should be controlled by a placebo group C, for instance, because of safety concerns of the treatments, we can continue the three-armed trial even in the case of an early stage i_0 yielding significance in (13) without the risk of losing the already shown significance.

If we are satisfied with showing T as being noninferior to R, we will stop the trial after that stage j^*, when $Z_{j^*}^{TR}(-\Delta_o) > cv_{j^*}(TR)$ the first time and $j^* \ge i_0$.

In case unexpected, for T favourable estimates of the involved parameters in the groups T and R have been observed up to stage $j^* < K$, this may lead to considerations to switch from showing noninferiority to showing superiority of T with respect to R. The trial is then continued by further planning with $\Delta = 0$ for the testing problem (2). Note that by (14), there is no risk to lose the noninferiority once shown.

2.2 Homogeneity of the Treatment Differences

Let us consider the extended model that each stage has individual parameters, say $\mu_{d,i}$ and $\sigma_{d,i}^2 > 0$, $d = T, R, C$, $i = 1, \ldots, K$. Then

$$\theta_{TC,i} = \mu_{T,i} - \mu_{C,i} \quad \text{and} \quad \theta_{TR,i} = \mu_{T,i} - \mu_{R,i}, \quad i = 1, \ldots, K, \qquad (15)$$

are the stage specific parameters for the treatment differences. The distributions of the test statistics $D_i^{TC}(\theta_{TC,i})$, see (4), and $D_i^{TR}(\theta_{TR,i})$, see (5), remain valid, so that, see (9),

$$z_i^d(\theta_{d,i}) \sim \mathcal{N}(0,1), \quad d = TC, TR, \ i = 1, \ldots, K. \qquad (16)$$

So as in (10), we get

$$\widetilde{Z}_j^d(\theta_{d,1},\ldots,\theta_{d,j}) := \sum_{i=1}^{j} z_i^d(\theta_{d,i}) \sim \sqrt{j}\,\mathcal{N}(0,1), \quad d = TC, TR,\ j = 1,\ldots,K, \tag{17}$$

and we can apply (11) with the positive critical values $cv_j(d)$ to give the following probability statement, see (12),

$$\begin{aligned} P_{\theta_{d,1},\ldots,\theta_{d,k}}\Big(-cv_j(d) \le \widetilde{Z}_j^d(\theta_{d,1},\ldots,\theta_{d,j}) \le cv_j(d) \text{ for } j = 1,\ldots,k \le K\Big) \\ \ge 1 - 2\alpha, \quad d = TC, TR. \end{aligned} \tag{18}$$

For example, with $d = TC$, the stage specific estimators $\hat{\theta}_{TC,i} = \bar{X}_{T_i} - \bar{X}_{C_i}$, $i = 1,\ldots,j \le k \le K$, satisfy the inequalities in the brackets of (18) because of $z_i^{TC}(\hat{\theta}_{TC,i}) = 0$.

When we assume that the parameters $\theta_{d,i}$ are really identical up to the k-th stage, say $\theta_{d,i} = \theta_d$ for $i = 1,\ldots,k$, then $\widetilde{Z}_j^d(\theta_{d,1},\ldots,\theta_{d,j}) = Z_j^d(\theta_d)$ from (10), $j = 1,\ldots,k$, and by (18), there holds

$$\begin{aligned} P_{\theta_d}(-cv_j(d) \le Z_j^d(\theta_d) \le cv_j(d) \text{ for } j = 1,\ldots,k \le K) \ge 1 - 2\alpha, \\ d = TC, TR. \end{aligned} \tag{19}$$

If now for a common level α, we cannot find some value for θ_d satisfying the inequalities in (19), we can conclude with an error rate of at most 2α, that the assumption of identical parameters up to the k-th stage was wrong. This can formally be stated as a test on homogeneity of the treatment differences.

In testing, for $d = TC$ or TR,

$$H_{0,hom}^d(k) : \theta_{d,1} = \ldots = \theta_{d,k} \quad \text{versus} \quad H_{1,hom}^d(k) : \theta_{d,i_1} \ne \theta_{d,i_2} \tag{20}$$

for some $i_1, i_2 \in \{1,\ldots,k\}, k = 2,\ldots,K$, the homogeneity hypothesis $H_{0,hom}^d(k)$ will be rejected at level of at most 2α, if

$$\widetilde{CI}_k^d := \{y \in I\!R \mid -cv_j(d) \le Z_j^d(y) \le cv_j(d) \text{ for } j = 1,\ldots,k\} = \emptyset. \tag{21}$$

If $H_{0,hom}^d(k^*)$ is rejected, then also $H_{0,hom}^d(k)$ will be rejected for $k^* \le k \le K$. An alternative to this homogeneity test does not seem to be known. A possible way to verify (21) numerically will be provided in Section 3.2.

A specific group sequential homogeneity test was claimed, for instance, by Koch (2006), who also pointed out its need for regulatory concerns.

Suppose that up to stage $k-1$ the sets $\widetilde{CI}_j^d$, $j = 1, \ldots, k-1$, are nonempty and $\widetilde{CI}_k$ is empty for a common level α. Thus, up to an error rate of 2α, see (20), we may consider that the parameters of the underlying treatment difference 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. Nevertheless, one may try to find an explanation for the revealed treatment behavior.

The basic test principle applied here to (20) is the same as used by Hartung and Knapp (2003) in deriving a test on homogeneity of variances of random treatment-by-sample interactions. As usually done, the significance level in the homogeneity test may be chosen higher than in the efficiency test. In the extreme case, when α comes near $1/2$ in (19), all stage specific estimates of the treatment differences had to be nearly equal in order to avoid a rejection of the homogeneity hypothesis $H_{0,hom}$ in (20), at level near 1.

Additionally to (20), in an approximate way, homogeneity tests from meta-analysis may be applied, see Hartung, Knapp, and Sinha (2008, Chapter 6).

3 Group Sequential Confidence Intervals

The functions $F_{t_\nu}(\cdot)$ and $\Phi^{-1}(\cdot)$, used in (8) and (9), are (strictly) monotone increasing in their arguments. The pivotal test statistics $D_i^{TC}(\theta_{TC})$ and $D_i^{TR}(\theta_{TR})$ from (4) and (5) are monotone decreasing in θ_{TC} and θ_{TR}, respectively, implying that $z_i^d(\theta_d) = \Phi^{-1}(F_{t_{\nu_i(d)}}(D_i^d(\theta_d)))$, see (9), is monotone decreasing in θ_d, $d = TC, TR$, $i = 1, \ldots, k$. Consequently, we can state for the whole functions from (10):

$$\begin{gathered} Z_j^{TC}(\theta_{TC}) \text{ and } Z_j^{TR}(\theta_{TR}) \text{ are monotone decreasing in } \theta_{TC} \text{ and } \theta_{TR}, \\ \text{respectively, } j = 1, \ldots, K. \end{gathered} \tag{22}$$

3.1 One-sided Confidence Intervals

From (21), we derive the lower confidence sets on θ_d as

$$\begin{gathered} CI_{k,\mathrm{I}}^d(\theta_d) := \{y \in I\!R \mid Z_j^d(y) \leq cv_j(d) \text{ for } j = 1, \ldots, k\}, \\ d = TC, TR,\ k = 1, \ldots, K, \end{gathered} \tag{23}$$

and again by (12), the confidence coefficient of $CI_{K,\mathrm{I}}^d$ is at least $1-\alpha$ and exactly $1-\alpha$ for $k = K$. Further, the confidence sets are nested,

$$CI_{k+1,\mathrm{I}}^d(\theta_d) \subset CI_{k,\mathrm{I}}^d(\theta_d), \quad k = 1, \ldots, K-1,\ d = TC, TR, \tag{24}$$

Using now that $Z_j^d(y)$ is monotone decreasing in y, see (22), we obtain that the confidence sets are genuine intervals, allowing the following representation:

$$CI_{k,\mathrm{I}}^d(\theta_d) = \left[L_k^d, \infty\right), \quad d = TC \text{ or } TR, \tag{25}$$

where $L_k^d = \max\{L^d(1), \ldots, L^d(k)\}$ and $L^d(j)$ solves

$$Z_j^d(L^d(j)) = cv_j(d), \quad j = 1, \ldots, k, \ k = 1, \ldots, K. \tag{26}$$

Note that the solutions $L^d(j)$ in (26) are unique and can be iteratively found, for instance, by use of the bisection method. Let us apply the group sequential confidence intervals to our hierarchical testing problem at overall significance level α, $0 < \alpha < 1/2$. Since the intervals are nested, see (24), we obtain in accordance with (13) and (14) by use of L_k^d from (25) the following decision rules:

At stage k, $k = 1, \ldots, K$,

$$\begin{aligned}
&\text{(i) if } 0 \geq L_k^{TC}, \text{ then stay with } H_0^{TC} \text{ in (1)} \\
&\quad \text{and } H_{0,\Delta_0}^{TR} \text{ in (2)}, \\
&\text{(ii) if } 0 < L_k^{TC} \text{ and } -\Delta_0 \geq L_k^{TR}, \text{ then decide for } H_1^{TC} \text{ in (1)} \\
&\quad \text{and stay with } H_{0,\Delta_0}^{TR} \text{ in (2)}, \\
&\text{(iii) if } 0 < L_k^{TC} \text{ and } -\Delta < L_k^{TR}, \text{ then decide for } H_1^{TC} \text{ in (1)} \\
&\quad \text{and for } H_{1,\Delta}^{TR} \text{ in (2)}, \ \Delta \in [0, \Delta_0].
\end{aligned} \tag{27}$$

In case (iii), we may stop the trial after stage k. If at some stage $j^* < K$ we observe $-\Delta_0 < L_{j^*}^{TR}$, we may consider, when continuing the study, to switch in the further planning to show T as superior to R.

In case (ii), if $k < K$, we may stop the control arm C and continue the trial only with the arms T and R.

In case (i), we have to continue the trial provided $k < K$.

3.2 Two-sided Confidence Intervals

In analogy to (23), let us define the upper confidence sets on θ_d as

$$\begin{gathered}
CI_{k,\mathrm{II}}^d(\theta_d) := \{y \in I\!R \mid -cv_j(d) \leq Z_j^d(y) \text{ for } j = 1, \ldots, k\}, \\
d = TC, TR, \ k = 1, \ldots, K,
\end{gathered} \tag{28}$$

and by (12), each confidence set has a confidence coefficient of at least $1-\alpha$, being exactly $1-\alpha$ for $k=K$. The interval representation, using (22), is given by

$$CI^d_{k,\mathrm{II}}(\theta_d) = \left(-\infty, U^d_k\right], \quad d = TC \text{ or } TR, \tag{29}$$

where $U^d_k = \min\{U^d(1), \ldots, U^d(k)\}$ and $U^d(j)$ solves uniquely

$$Z^d_j(U^d(j)) = -cv_j(d), \quad j = 1, \ldots, k, \ k = 1, \ldots, K. \tag{30}$$

The two-sided confidence interval on θ_d at stage k is defined as the intersection of the two corresponding one-sided confidence intervals,

$$CI^d_k(\theta_d) := \left[L^d_k, U^d_k\right], \quad d = TC \text{ or } TR, \tag{31}$$

where L^d_k is from (25) and U^d_k is from (29), $k = 1, \ldots, K$. The confidence intervals are nested,

$$CI^d_{k+1}(\theta_d) \subset CI^d_k(\theta_d), \quad k = 1, \ldots, K-1, \ d = TC, TR, \tag{32}$$

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

Denote $I^d_k(\theta_d) = [L^d(j), U^d(j)]$, see (26) and (30), the individual two-sided confidence interval on θ_d at the k-stage. Then it holds,

$$\begin{aligned} &CI^d_1(\theta_d) = I^d_1(\theta_d) \text{ and} \\ &CI^d_k(\theta_d) = CI^d_{k-1}(\theta_d) \cap I^d_k(\theta_d), \quad k = 2, \ldots, K, \ d = TC, TR, \end{aligned} \tag{33}$$

Since $CI^d_k \subset I^d_k$, the interval I^d_k is another two-sided confidence interval with confidence coefficient of at least $1-2\alpha$. The interval I^d_k results from the boundaries in stage k alone and will be always nonempty. Therefore, I^d_k may be preferred to CI^d_k, see for instance Jennison and Turnbull (2000, p. 192) in their corresponding setting. Depending on the choice of α, the two-sided confidence interval $CI^d_k(\theta_d)$ from (31) may be empty.

Let us look at the homogeneity test (20). We have to check whether the set $\widetilde{CI}^d_k$ defined in (21) is empty. Now we can state that this set coincides with the two-sided confidence interval: $\widetilde{CI}^d_k = CI^d_k(\theta_d)$. Hence, if $CI^d_k(\theta_d)$ is empty, that is $U^d_k < L^d_k$, see (31), we have to reject the homogeneity hypothesis $H^d_{0,hom}(k)$ in (20) with an error rate of at most 2α. Consequently, preferring I^d_k to CI^d_k does not provide some real advantage.

On the other hand, under the model assumptions of identical parameters underlying the different stages of the study, the probability to obtain an empty confidence interval $CI^d_k(\theta_d)$ is bounded by 2α, $d = TC$ or TR.

3.3 Approximative Confidence Intervals

Instead of the implicitly defined confidence intervals, we provide approximative confidence intervals in an explicit form. Their boundaries may be used also as starting points in an iterative procedure to determine the exact confidence intervals.

Let us approximate the central t-distributions involved in the combining statistics by normal distributions with the same first two moments. The variance of a t_ν-variate is $\nu/(\nu-2)$. So we may define the following weights at the i-th stage, $i = 1, \ldots, K$,

$$w_i^{TC} := \sqrt{\frac{\nu_i(TC) - 2}{\nu_i(TC)[\hat{\sigma}_{T_i}^2/n_{T_i} + \hat{\sigma}_{C_i}^2/n_{C_i}]}}, \tag{34}$$

provided $\nu_i(TC) > 2$, see (4), and thus, the statistic $z_i^{TC}(\theta_{TC})$ from (9) is approximated by

$$z_i^{TC}(\theta_{TC})_{appr} = \Phi^{-1}\left(\Phi\left[w_i^{TC}(\bar{X}_{T_i} - \bar{X}_{C_i} - \theta_{TC})\right]\right), \tag{35}$$

which is approximately $\mathcal{N}(0,1)$-distributed. Hence, the combining statistic $Z_j^{TC}(\theta_{TC})$ from (10) is approximated by

$$Z_j^{TC}(\theta_{TC})_{appr} = \sum_{i=1}^{j} w_i^{TC}(\bar{X}_{T_i} - \bar{X}_{C_i} - \theta_{TC}), \quad j = 1, \ldots, K, \tag{36}$$

which is approximately $\mathcal{N}(0, j)$-distributed. Equating $Z_j^{TC}(y)_{appr}$ to $cv_j(TC)$, see (26), and to $-cv_j(TC)$, see (30), and solving for y yields the following approximate individual confidence interval on θ_{TC}, see (33), for $j = 1, \ldots, K$,

$$I_j^{TC}(\theta_{TC})_{appr} = \sum_{i=1}^{j} \frac{w_i^{TC}(\bar{X}_{T_i} - \bar{X}_{C_i})}{\sum_{h=1}^{j} w_h^{TC}} \pm \frac{cv_j(TC)}{\sum_{h=1}^{j} w_h^{TC}}. \tag{37}$$

By setting

$$\begin{aligned} CI_1^{TC}{}_{appr} &= I_1^{TC}(\theta_{TC})_{appr} \text{ and} \\ CI_k^{TC}(\theta_{TC})_{appr} &= CI_{k-1}^{TC}(\theta_{TC})_{appr} \cap I_k^{TC}(\theta_{TC})_{appr}, \quad k = 2, \ldots, K, \end{aligned} \tag{38}$$

we obtain approximations of the confidence intervals CI_k^{TC} on $\theta_{TC} = \mu_T - \mu_C$ in (31). Proceeding analogously, we get approximate confidence intervals on $\theta_{TR} = \mu_T - \mu_R$.

3.4 Confidence Intervals on the Means

Let, based on the data of the i-th stage, $i = 1, \ldots, K$, $\hat{\sigma}_i^2(T)$ be an unbiased estimator of σ_T^2, which is stochastically independent of $\bar{X}_{T_i}$ and satisfies, see (3), (6) or (7),

$$\nu_i(T)\ \hat{\sigma}_i^2(T)/\sigma_T^2 \sim \chi_{\nu_i(T)}^2, \tag{39}$$

then with the i-th test statistic

$$D_i^T(\mu_T) := \sqrt{n_{T_i}}\,\frac{\bar{X}_{T_i} - \mu_T}{\sqrt{\hat{\sigma}_i^2(T)}} \sim t_{\nu_i}(T), \quad i = 1, \ldots K, \tag{40}$$

we derive the combining statistics, see (10), for $j = 1, \ldots, K$,

$$Z_j^T(\mu_T) := \sum_{i=1}^{j} \Phi^{-1}\left(F_{t_{\nu_i(T)}}\left(D_i^T(\mu_T)\right)\right) \sim \sqrt{j}\,\mathcal{N}(0,1). \tag{41}$$

In the same way as above, see (31), we obtain the two-sided confidence interval on μ_T at the k-th stage, $k = 1, \ldots, K$, as

$$CI_k^T(\mu_T) = \left[\max_{1\le j\le k} L^T(j), \min_{1\le j\le k} U^T(j)\right], \tag{42}$$

where with positive critical values $cv_j(T)$ in (11), $L^T(j)$ and $U^T(j)$ are the unique solutions of

$$Z_j^T(L^T(j)) = cv_j(T) \text{ and } Z_j^T(U^T(j)) = -cv_j(T), \quad j = 1, \ldots, k \le K. \tag{43}$$

The confidence intervals are nested and possess confidence coefficients of at least $1 - 2\alpha, 0 < \alpha < 1/2$.

Defining the weights, for $i = 1, \ldots, K$,

$$w_i^T := \sqrt{\frac{(\nu_i(T) - 2)n_{T_i}}{\nu_i(T)\hat{\sigma}_i^2(T)}}, \quad \nu_i(T) > 2, \tag{44}$$

we receive the approximate individual confidence interval on μ_T at the j-th stage, see (37), for $i = 1, \ldots, K$,

$$I_j^T(\mu_T)_{appr} := \sum_{i=1}^{j} \frac{w_i^T \bar{X}_{T_i}}{\sum_{h=1}^{j} w_h^T} \pm \frac{cv_j(T)}{\sum_{h=1}^{j} w_h^T}, \tag{45}$$

Again by $CI_1^T(\mu_T)_{appr} = I_1^T(\mu_T)_{appr}$ and $CI_k^T(\mu_T)_{appr} = CI_{k-1}^T(\mu_T)_{appr} \cap I_k^T(\mu_T)_{appr}, k = 2, \ldots, K$, we obtain approximations of $CI_k^T(\mu_T)$ in (42). Confidence intervals on μ_C and μ_R are derived in the same way.

3.5 Confidence Intervals on the Variance Parameters

Let $F_{\chi_\nu^2}$ denote the cumulative distribution function of a χ^2-variate with ν degrees of freedom. Using the pivotal χ^2-statistics from (39), which are monotone decreasing in $\sigma_T^2 > 0$, we obtain, in analogy to (8),

$$F_{\chi^2_{\nu_i(T)}}\left(\nu_i(T)\,\frac{\hat{\sigma}_i^2(T)}{\sigma_T^2}\right) \sim \mathcal{U}(0,1), \quad i = 1, \ldots, K, \tag{46}$$

leading as in (10) to the pivotal combining statistics, for $j = 1, \ldots, K$,

$$V_j^T(\sigma_T^2) := \sum_{i=1}^{j} \Phi^{-1}\left[F_{\chi^2_{\nu_i(T)}}\left(\nu_i(T)\ \frac{\hat{\sigma}_i^2(T)}{\sigma_T^2}\right)\right] \sim \sqrt{j}\ \mathcal{N}(0,1), \tag{47}$$

which are monotone decreasing in $\sigma_T^2 > 0$. Denote $cv_1^*, \ldots, cv_K^*$ positive critical values defined by (11).

Let $\sigma_{T,L}^2(j)$ and $\sigma_{T,U}^2(j)$ be the unique solutions of the equations

$$V_j^T(\sigma_{T,L}^2(j)) = cv_j^* \text{ and } V_j^T(\sigma_{T,U}^2(j)) = -cv_j^*, \quad j = 1, \ldots, k \leq K, \tag{48}$$

then in analogy to (31), we derive the confidence intervals on σ_T^2 as

$$VCI_k^T(\sigma_T^2) = \left[\max_{1\leq j\leq k} \sigma_{T,L}^2(j), \min_{1\leq j\leq k} \sigma_{T,U}^2(j)\right], \quad k = 1, \ldots, K, \tag{49}$$

which are nested and possess confidence coefficients of at least $1 - 2\alpha, 0 < \alpha < 1/2$. For common α, an empty interval indicates that the assumption of homogeneous variances over the stages may be violated, see (20) and Section 3.2.

Applying the rule of error propagation, we derive for the χ^2-statistics of (39), that the transformations

$$g_i^T(\sigma_T^2) := \sqrt{2\ \frac{\nu_i(T)\ \hat{\sigma}_i^2(T)}{\sigma_T^2}} - \sqrt{2\ \nu_i(T)}, \quad i = 1, \ldots, K, \tag{50}$$

are approximately $\mathcal{N}(0,1)$-distributed, so that $G_j^T(\sigma_T^2) := \sum_{i=1}^{j} g_i^T(\sigma_T^2)$ is approximately $\mathcal{N}(0,j)$-distributed, $j = 1, \ldots, K$. Solving now $G_j^T(y) \leq cv_j^*$ and $G_j^T(y) \geq -cv_j^*$ for $y > 0$ yields the approximate individual confidence intervals, see (33), on σ_T^2 as

$$VI_j^T(\sigma_T^2)_{appr} := [a_T(j)^2, b_T^*(j)^2], \quad j = 1, \ldots, K, \tag{51}$$

where

$$b_T^*(j)^2 = \begin{cases} b_T(j)^2 & \text{if } b_T(j) > 0, \\ \infty & \text{if } b_T(j) \leq 0, \end{cases}$$

$$a_T(j) = \frac{\sum_{i=1}^{j} \sqrt{2\ \nu_i(T)}\ \hat{\sigma}_i(T)}{\sum_{h=1}^{j} \sqrt{2\ \nu_h(T)} + cv_j^*} \text{ and}$$

$$b_T(j) = \frac{\sum_{i=1}^{j} \sqrt{2\ \nu_i(T)}\ \hat{\sigma}_i(T)}{\sum_{h=1}^{j} \sqrt{2\ \nu_h(T)} - cv_j^*}.$$

Again by setting $VCI_1^T(\sigma_T^2)_{appr} = VI_1^T(\sigma_T^2)_{appr}$ and $VCI_k^T(\sigma_T^2)_{appr} = VCI_{k-1}^T(\sigma_T^2)_{appr} \cap VI_k^T(\sigma_T^2)_{appr}$, $k = 2, \ldots, K$, we get explicit approximations to the confidence intervals on σ_T^2 in (49). In the same way, we may derive confidence intervals on σ_C^2 and σ_R^2.

4 Group Sequential Point Estimation

4.1 Estimation of the Treatment Difference

For $\theta_{TC} = \mu_T - \mu_C$, the combining statistic $Z_j^{TC}(\theta_{TC})$ from (10) is $\mathcal{N}(0, j)$-distributed with mode and median 0. The *maximum likelihood (ML) estimator* $\hat{\theta}_{TC}^{(1)}(j)$ of θ_{TC} at stage j is given by

$$\hat{\theta}_{TC}^{(1)}(j) \text{ solves } Z_j^{TC}\left(\hat{\theta}_{TC}^{(1)}(j)\right) = 0, \quad j = 1, \ldots, K. \tag{52}$$

The solution in (52) is unique.

The *global* p-value at stage j is

$$p_{TC}(j) = 1 - \Phi\left(Z_j^{TC}(\theta_{TC})/\sqrt{j}\right), \quad j = 1, \ldots, K, \tag{53}$$

and solving (53) for θ_{TC} such that $p_{TC}(j) = 1/2$ yields $\hat{\theta}_{TC}^{(1)}(j)$ as solution. Since $Z_j^{TC}(\theta)$ is monotone in θ_{TC}, we can conclude:

$$\hat{\theta}_{TC}^{(1)}(j) \text{ is } \textit{median unbiased}, \quad j = 1, \ldots, K, \tag{54}$$

see Cox and Hinkley (1974, p. 273), that is, the ML-estimator $\hat{\theta}_{TC}^{(1)}(j)$ lies with equal probability as well below the parameter θ_{TC} as above θ_{TC}.

Equating the approximative combining statistic $Z_j^{TC}(\theta_{TC})_{\text{appr}}$ from (36) to 0 and solving for θ_{TC} yields the midpoint of the approximative individual confidence interval $I_j^{TC}(\theta_{TC})_{\text{appr}}$ from (37) as *approximate* median unbiased ML-estimator $\hat{\theta}_{TC}^{(2)}(j)$ of θ_{TC} at the j-th stage, given by

$$\hat{\theta}_{TC}^{(2)}(j) = \sum_{i=1}^{j} \frac{w_i^{TC}(\bar{X}_{T_i} - \bar{X}_{C_i})}{\sum_{h=1}^{j} w_h^{TC}}, \quad j = 1, \ldots, K, \tag{55}$$

where the weights are defined in (34). Note that, in combining the mean differences of the stages, their inverse estimated standard errors are used in the weights and not their inverse estimated variances as known from the 'minimum variance unbiased' estimator of the overall mean difference in meta-analysis, see Hartung, Knapp, and Sinha (2008, Chapter 8). Weighted mean differences like $\hat{\theta}_{TC}^{(2)}(j)$ from (55) are used in the generalized Cochran-Wald statistics considered by Hartung, Böckenhoff, and Knapp (2003).

Replacing in (55) the weights w_i^{TC} by

$$\tilde{w}_i^{TC} = \left(\frac{\hat{\sigma}_{T_i}^2}{n_{T_i}} + \frac{\hat{\sigma}_{C_i}^2}{n_{C_i}}\right)^{-1}, \quad i = 1, \ldots, K, \tag{56}$$

we obtain the *meta-analytical* estimator $\hat{\theta}_{TC}^{(3)}(j)$ of θ_{TC} up to the j-th stage, $j = 1, \ldots, K$. For $\theta_{TR} = \mu_T - \mu_R$, the estimators $\hat{\theta}_{TR}^{(h)}(j)$ of θ_{TR} at stage j, $h = 1, 2, 3$, are defined analogously.

4.2 Estimation of the Mean and Variance Parameters

With the combining statistic from (41), the unique solution of $Z_j^T\left(\hat{\mu}_T^{(1)}(j)\right) = 0$ defines the *median unbiased ML-estimator* $\hat{\mu}_T^{(1)}(j)$ of μ_T at the j-th stage, $j = 1, \ldots, K$, see above. The midpoint of the interval $I_j^T(\mu_T)_{\text{appr}}$ in (45) is the *approximate* median unbiased ML-estimator $\hat{\mu}_T^{(2)}(j)$ of μ_T at the j-th stage, and the replacing there the weight w_i^T by $\tilde{w}_i^T = n_{T_i}/\hat{\sigma}_i^2(T)$ yields the midpoints as the *meta-analytical* estimator $\hat{\mu}_T^{(3)}(j)$ of μ_T at the j-th stage, $j = 1, \ldots, K$.

By a quite analogous argumentation as above in Section 4.1, we derive with the combining statistic from (47) the *median unbiased ML-estimator* $\widehat{\sigma_T^2}^{(1)}(j)$ of σ_T^2 at the j-th stage as follows:

$$\widehat{\sigma_T^2}^{(1)}(j) \text{ solves uniquely } V_j^T\left(\widehat{\sigma_T^2}^{(1)}(j)\right) = 0, \quad j = 1, \ldots, K. \tag{57}$$

The solution of $G_j^T(y) = 0$, see (50), (51), is the *approximate median* unbiased ML-estimator of σ_T^2 at the j-th stage, which can be represented as:

$$\widehat{\sigma_T^2}^{(2)}(j) = \left(\sum_{i=1}^{j} \frac{\sqrt{\nu_i(T)}\ \hat{\sigma}_i(T)}{\sum_{h=1}^{j}\sqrt{\nu_h(T)}}\right)^2, \quad j = 1, \ldots, K. \tag{58}$$

Since the variance of $\hat{\sigma}_i^2(T)$ from (39) is $2\sigma_T^4/\nu_i(T)$, the *meta-analytical* inverse variance weighted estimator of σ_T^2 up to the j-th stage takes on the following form:

$$\widehat{\sigma_T^2}^{(3)}(j) = \sum_{i=1}^{j} \frac{\nu_i(T)\ \hat{\sigma}_i^2(T)}{\sum_{h=1}^{j}\nu_h(T)}, \quad j = 1, \ldots, K, \tag{59}$$

which may also be considered at the *pooled* estimator of σ_T^2 up to the j-th stage. Note that in (58) the estimated standard deviations are combined, and in (59), the estimated variances. In the groups R and C, the parameters are estimated in an analogous way.

5 Sample Size Calculation and Adaptive Updating

Suppressing the subscript i and supposing known variances, let us consider the test statistic, see (5),

$$D_0^{TR}(\theta_{TR}) = \frac{\bar{X}_T - \bar{X}_R - \theta_{TR}}{\sqrt{\sigma_T^2/n_T + \sigma_R^2/n_R}} \sim \mathcal{N}(0,1), \tag{60}$$

which should be used for testing, with fixed $\Delta \in [0, \Delta_0]$ and fixed value $\theta_{TR}^* > -\Delta$, the point hypotheses

$$H_0^* : \ \theta_{TR} = -\Delta \quad \text{versus} \quad H_1^* : \ \theta_{TR} = \theta_{TR}^* > -\Delta, \tag{61}$$

so that under H_0^*, $D_0^{TR}(-\Delta) \sim \mathcal{N}(0,1)$.

Then for given level α, $0 < \alpha < 1$, and desired power $1 - \beta_{TR}$, $0 < \beta_{TR} < 1$, the required sample sizes n_T and n_R should satisfy the following inequality,

$$\frac{\theta_{TR}^* - (-\Delta)}{\sqrt{\sigma_T^2/n_T + \sigma_R^2/n_R}} \geq \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta_{TR}). \tag{62}$$

For ease of presentation, we will use this formula (62) as a good approximation also in the following, when t-statistics will be applied. Denote stage 0 a priori information and external restrictions.

Then after stage j, based on previous information of stages $0, 1, \ldots, j$, let $\hat{\theta}_{TR}(j) > -\Delta$, $\hat{\theta}_{TC}(j) > 0$, $\hat{\sigma}_T^2(j)$, $\hat{\sigma}_R^2(j)$, and $\hat{\sigma}_C^2(j)$, $j = 0, 1, \ldots, K$, be reasonable estimates of their corresponding parameters, for instance, by use of the point estimators provided in Section 4. Assume the test above is placed after stage j, and put $\theta_{TR}^* = \hat{\theta}_{TR}(j)$, $\sigma_T^2 = \widehat{\sigma_T^2}(j)$, and $\sigma_R^2 = \widehat{\sigma_R^2}(j)$, then formula (62) becomes the following inequality, with $\hat{\theta}_{TR}(j) + \Delta > 0$,

$$\frac{\hat{\theta}_{TR}(j) + \Delta}{\sqrt{\widehat{\sigma_T^2}(j)/n_T + \widehat{\sigma_R^2}(j)/n_R}} \geq \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta_{TR}), \quad j = 0, \ldots, K. \tag{63}$$

Whereas $1 - \beta_{TR}$ is the desired power at $\theta_{TR}(j) = \hat{\theta}_{TR}(j) > -\Delta$ in the testing problem (2) after stage j, let the desired power at $\theta_{TC} = \hat{\theta}_{TC}(j) > 0$ in the testing problem (1) after stage j be $1 - \beta_{TC}$, $0 < \beta_{TC} < 1$. So, with the same level α and by use of $D_0^{TC}(\theta_{TC})$, see (60), we derive analogously for the required sample sizes n_T and n_C in the testing problem (1) after stage j the following inequality, with $\hat{\theta}_{TC}(j) > 0$,

$$\frac{\widehat{\theta}_{TC}(j)}{\sqrt{\widehat{\sigma_T^2}(j)/n_T + \widehat{\sigma_C^2}(j)/n_C}} \geq \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta_{TC}), \quad j = 0, \ldots, K. \tag{64}$$

For an easy use of these formulae in the following, let us define sets of feasible sample sizes, for $k = 0, \ldots, K$,

$$\Gamma_{TR}(\kappa, \beta_{TR}, \Delta)_k := \{(n_T, n_R) \in I\!N \times I\!N | n_T \text{ and } n_R \text{ satisfy (63) for } j = k \text{ and } \alpha = \kappa\} \quad (65)$$

$$\Gamma_{TC}(\kappa, \beta_{TC})_k := \{(n_T, n_C) \in I\!N \times I\!N | n_T \text{ and } n_C \text{ satisfy (64) for } j = k \text{ and } \alpha = \kappa\}\,. \quad (66)$$

Recall now from (11), for $d = TC$ or TR, the event

$$\mathrm{A} := \left\{ \sum_{i=1}^{h} Y_i \leq c\nu_h(d) \text{ for all } h = 1, \ldots, K \right\},$$

and let us consider for an arbitrary, but fixed, stage j, $j \in \{1, \ldots, K\}$, the event

$$\mathrm{B} := \left\{ \sum_{i=1}^{h} Y_i \leq c\nu_h \text{ for } h = 1, \ldots, j-1, \text{ and } \sum_{i=1}^{j-1} Y_i + \sum_{i=j}^{K} Y_i \leq c\nu_K(d) \right\}.$$

Clearly, the probability of event B is larger than of event A. Moreover, $\sum_{i=j}^{K} Y_i$ is $\mathcal{N}(0,\ K-(j-1))$-distributed and may be collapsed to $\sqrt{K-(j-1)}Y_j$, which has the same distribution. Hence, we obtain

$$\left\{ \sum_{i=1}^{h} Y_i \leq c\nu_h(d) \text{ for } h = 1, \ldots, j-1, \text{ and } \sum_{i=1}^{j-1} Y_i + \sqrt{K-(j-1)}Y_j \leq c\nu_K(d) \right\}$$
$$\supset \left\{ \sum_{i=1}^{h} Y_i \leq c\nu_h(d) \text{ for all } h = 1, \ldots, K \right\}. \quad (67)$$

Further, denote θ_d^0 a value for θ_d under the null-hypothesis H_0^d, given as H_0^{TC} from (1) for $d = TC$ or as $\mathrm{H}_{0,\Delta}^{TR}$ from (2) for $d = TR$ and $\Delta \in [0, \Delta_0]$ fixed. The aim is, that at some stage j, by use of the combining pivotal statistic from (10), we will obtain: $Z_j^d(\theta_d^0) > c\nu_j(d)$.

The following proceeding is a consequence of (67). 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, named stage (j, K), and build the final test statistic, see (9) and (10),

$$Z_{(j,K)}^d(\theta_d^0) = Z_{j-1}^d(\theta_d^0) + \sqrt{K-(j-1)}\ \Phi^{-1}\left[1 - p_{(j,K)}^d(\theta_d^0)\right], \quad (68)$$

where $Z_{(j,K)}^d(\theta_d^0) \sim \sqrt{K}\ \mathcal{N}(0,1)$ under H_0^d, $j = 1, \ldots, K$, defining $Z_0^d = 0$. The test statistic $Z_{(j,K)}^d(\theta_d^0)$ has to be compared with the K-th critical value $c\nu_K(d)$ in testing H_0^d. We want to reach $c\nu_K(d)$ by use of $Z_{(j,K)}^d(\theta_d^0)$, that is, we have to equate:

$$c\nu_K(d) = Z_{j-1}^d(\theta_d^0) + \sqrt{K-(j-1)}\ \Phi^{-1}\left[1 - p_{(j,K)}^d(\theta_d^0)\right], \quad (69)$$

and solving the equation for the unknown p-value yields as solution the projected p-value

$$\hat{p}^d_{(j,K)}(\theta^0_d) = 1 - \Phi\left[\frac{c\nu_K(d) - Z^d_{j-1}(\theta^0_d)}{\sqrt{K-(j-1)}}\right], d = TC \text{ or } TR, j = 1, \ldots, K. \tag{70}$$

Now $c\nu_K(d)$ will be reached with probability $1-\beta_d$ when the stage specific t-tests, concerning $d = TC$ or TR, which are planned for the next final stage (j,K), will provide levels attained or p-values below the projected p-values with probability $1-\beta_{TC}$ and $1-\beta_{TR}$, respectively. Observed p-values below the projected p-values would yield: $Z^d_{(j,K)}(\theta^0_d) > c\nu_K(d)$, $d = TC$ or TR. Thus, the significance or α-level of these tests, say $\alpha^d_{(j,K)}$, $d =$ TC, TR, are chosen to satisfy

$$\alpha^d_{(j,K)} = \hat{p}^d_{(j,K)}(\theta^0_d),\ d = TC, TR, \quad j = 1, \ldots, K. \tag{71}$$

Consequently, conditioned on $\theta_{TC} = \hat{\theta}_{TC}(j-1) > 0$ and $\theta_{TR} = \hat{\theta}_{TR}(j-1) > -\Delta$, when the above mentioned t-tests would be applied, the required sample sizes M_{T_j}, M_{C_j} and M_{R_j} of the respective groups in the final stage (j,K), for holding the power $1-\beta_{TC}$ for $d = TC$ in (1) and the power $1-\beta_{TR}$ for $d = TR$ in (2), should be feasible and satisfy:

$$(M_{T_j}, M_{C_j}) \in \Gamma_{TC}(\hat{p}^{TC}_{(j,K)}(0), \beta_{TC})_{j-1} \text{ and} \tag{72}$$

$$(M_{T_j}, M_{R_j}) \in \Gamma_{TR}(\hat{p}^{TR}_{(j,K)}(-\Delta), \beta_{TR}, \Delta)_{j-1}, \tag{73}$$

see (65), (66), (70), (71). Thus, the projected p-values can be named as conditional error functions.

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 will not perform the above mentioned t-tests in the final stage (j,K) but in stage j. Consequently, we will choose now the sample size in each group for stage j proportionally as, see (72) and (73),

$$n_{T_j} \approx \frac{M_{T_j}}{K-j+1}, \quad n_{C_j} \approx \frac{M_{C_j}}{K-j+1}, \quad n_{R_j} \approx \frac{M_{R_j}}{K-j+1}, \quad j = 1, \ldots, K, \tag{74}$$

which is a (slightly) conservative choice by (67). Note that each sample size should be at least 2 in each stage. Then we use $c\nu_j(TC)$ as critical value for $Z^{TC}_j(0)$ and $c\nu_j(TR)$ as critical value for $Z^{TR}_j(-\Delta)$, $\Delta \in [0, \Delta_0)$, in stage j, see (10).

Especially for $j = 1$:

$$n_{T_1} \approx \frac{M_{T_1}}{K},\ n_{C_1} \approx \frac{M_{C_1}}{K},\ \text{and } n_{R_1} \approx \frac{M_{R_1}}{K} \tag{75}$$

where, see (65) and (66),

$$(M_{T_1}, M_{C_1}) \in \Gamma_{TC}(\alpha_{TC}, \beta_{TC})_0, \quad \alpha_{TC} := 1 - \Phi(c\nu_K(TC)/\sqrt{K}),$$

$$(M_{T_1}, M_{R_1}) \in \Gamma_{TR}(\alpha_{TR}, \beta_{TR}, \Delta)_0, \quad \alpha_{TR} := 1 - \Phi(c\nu_K(TR)/\sqrt{K}),$$

are feasible *starting sample sizes*.

By taking the initial sample sizes in all stages, (75) provides *formulae for sample size calculation* in non-adaptive group sequential trials.

Further, formulae (72) to (75) provide an *optimal allocation of sample sizes* at each stage, when minimizing, for instance, the total sample size at each stage under some side conditions. Often in practice, a chosen randomization scheme of the treatments has to be taken into account.

We start with the above calculated initial sample sizes in the first stage of the study. Then with the proceeding above, we reach the full power $1 - \beta_{TC}$, conditioned on $\theta_{TC} = \hat{\theta}_{TC}(K-1) > 0$ and $1 - \beta_{TR}$, conditioned on $\theta_{TR} = \hat{\theta}_{TR}(K-1) > -\Delta$, latest in stage $j = K$, if not stopped before because of shown significance.

The *total power*, say $1 - \beta_{\text{Total}}$, of the hierarchical testing of (1) and (2), can be estimated by

$$1 - \beta_{TC} - \beta_{TR} \leq 1 - \beta_{\text{Total}} \leq \min\{1 - \beta_{TC}, 1 - \beta_{TR}\}. \tag{76}$$

When we replace the combining statistic Z_j by its approximation $Z_{j,\text{appr}}$ from (36) in the above considerations, we obtain an approximative proceeding, that fulfills the purpose of sample size calculation in practical situations.

Further, we may formally define the p-values, see (8), as suiting to the null-hypothesis that θ_d 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 chosen adaptively in a consecutive way, see for instance Brannath, Posch, and Bauer (2002). All the above procedures are based on such p-values. Consequently, all the statements remain valid when sample sizes are chosen adaptively as demonstrated in this section, see also Hartung (2006).

Table 1: Controlled noninferiority clinical trial concerning patients with asthma bronchiale in an adaptive 3-stage Pocock (1977) design with early stop for shown significance after stage 2 at given one-sided significance level $\alpha = 0.025$.

Stage	Adaptive sample size			Data [in ℓ]				Test value			Critical value
i	n_{T_i}	n_{R_i}	n_{C_i}	$\bar{x}_{T_i}$	$\bar{x}_{R_i}$	$\bar{x}_{C_i}$	$\hat{\sigma}_i$	$Z_i^{TC}(0)$	$Z_i^{TR}(-0.2)$	$Z_i^{TR}(0)$	$c\nu_i$
0	—			2.6	2.5	2.1	0.9	—	$\Delta_0 = 0.2$	—	
1	116	58	29	2.65	2.56	2.13	0.87	2.86	2.06	0.45	2.289
2	96	48	24	2.69	2.51	2.15	0.81	5.76	4.70	1.71	3.237
3	STOP Because of shown significance										

6 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 test drug T and a standard reference drug R, for treating patients 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). A control group C received a placebo drug, but the same basic treatment as the groups T and R. Methods of the previous sections will be demonstrated for the hierarchical testing problem (1) and (2) in the present application, where $\Delta_0 = 0.2\ell$ is a usual margin of the noninferiority parameter for the considered clinical variable. To increase the number of observations on the new drug, the three drugs were randomized in blocks of size 7 containing each: 4×T, 2×R, 1×C. By technical reasons, the randomization scheme could not be changed in an interim analysis. Further, a common variance σ^2 could be assumed in all groups so that formulae (7) is used for its estimation and computing the degrees of freedom at each stage.

The one-sided significance level is chosen as $\alpha = 0.025$, the power in showing T as superior to C should be $1-\beta_{TC} = 0.95$, and $1-\beta_{TR} = 0.90$ should be the power for showing T as a noninferior to R, which means by (76) for the total power $1-\beta_{\text{Total}}$ of the hierarchical testing: $0.85 \leq 1 - \beta_{\text{Total}} \leq 0.90$. The study design was chosen as an adaptive three-stage design of Pocock (1977) type for both comparisons (1) and (2). Using the combining statistic from (10), we obtain in (11) the critical values $c\nu_j = 2.289\sqrt{j},\ j = 1, 2, 3$, see Hartung (2006, p. 533), or Jennsion and Turnbull (2000, p. 26) for the two-sided level 0.05.

Using the prior guesses of the parameters at stage 0 from Table 1, the differences $\theta_{TR} = \mu_T - \mu_R$ and $\theta_{TC} = \mu_T - \mu_C$ are initially estimated by $\hat{\theta}_{TR}(0) = 0.10\ell$ and $\hat{\theta}_{TC}(0) = 0.50\ell$, and σ by $\hat{\sigma}(0) = 0.90\ell$. Since the sample sizes have to satisfy $n_T = 2n_R$ and $n_T = 4n_C$, we get, by use of (75) from (72) and (73) the following conditions for M_{T_1}:

$$M_{T_1} \geq (1+4)(2.289+1.645)^2(0.9/0.5)^2 = 250.7, \text{ and}$$

$$M_{T_1} \geq (1+2)(2.289+1.282)^2(0.9/[0.1+0.2])^2 = 344.3.$$

Using (75), we compute a minimum of 29 blocks of size 7 for the first stage, such that $n_{T_1} = 116,\ n_{R_1} = 58,\ n_{C_1} = 29$.

The trial started with these numbers of patients. In the first stage we observed, see Table 1, $\bar{x}_{T_1} - \bar{x}_{C_1} = 0.52,\ \bar{x}_{T_1} - \bar{x}_{R_1} = 0.09$, and $\hat{\sigma}_1 = 0.87$, associated with $\nu_1 = 203 - 3$ degrees of freedom, see (7).

For the testing problem (1), we compute, see (10), $Z_1^{TC}(0) = 2.86 > 2.289 = c\nu_1$, such that by (13), the null-hypothesis H_0^{TC} can be rejected already after the first stage. For the testing problem (2), with $\Delta = \Delta_0 = 0.2$, we compute $Z_1^{TR}(-0.2) = 2.06$. In the further planning, we can look only on the testing in (2). By (70), we obtain the projected p-value

$$\hat{p}_{(2,3)}^{TR}(-0.2) = 1 - \Phi\left[\frac{2.289\sqrt{3} - 2.06}{\sqrt{2}}\right] = 1 - \Phi[1.3468].$$

Using the estimates of the first stage, that is, $\hat{\theta}_{TR}(1) = 0.09$ and $\hat{\sigma}_T^2(1) = \hat{\sigma}_R^2(1) = 0.87^2$, see (63), we obtain by (73),

$$M_{T_2} \geq (1+2)(1.3468+1.282)^2(0.87/[0.9+0.2])^2 = 186.6.$$

Using (74), we compute a minimum of 24 blocks of size 7 for the second stage, implying the sample sizes $n_{T_2} = 96,\ n_{R_2} = 48$ and $n_{C_2} = 24$.

In stage 2, we observed the estimates, see Table 1, $\bar{x}_{T_2} - \bar{x}_{R_2} = 0.18,\ \bar{x}_{T_2} - \bar{x}_{C_2} = 0.54$, and $\hat{\sigma}_2 = 0.81$, associated with $\nu_2 = 168 - 3$ degrees of freedom, see (7). We compute for the testing problem (2), $Z_2^{TR}(-0.2) = 2.06 + 2.64 = 4.70 > 2.289\sqrt{2} = 3.237 = c\nu_2$ such that the noninferiority of T with regard to R is shown, too. Consequently, the study is stopped after stage 2. The observed treatment effects seemed to be not favourable (or too expensive) for an attempt to reach superiority also in (2) at the third stage. Note that in the second stage, too, the test for (1) exceeds the critical value, $Z_2^{TC}(0) = 2.86 + 2.90 = 5.76 > 3.237$.

In the hierarchical testing problem, we could add as a third step to test at level α, too:

$$\mathrm{H}_0^{RC}: \ \mu_R = \mu_C \quad \text{versus} \quad \mathrm{H}_1^{RC}: \ \mu_R > \mu_C \tag{77}$$

provided both null-hypotheses H_0^{TC} in (1) and $\mathrm{H}_{0,\Delta_0}^{TR}$ in (2) are rejected at level α each. But then the sample sizes of both groups T and R should be equal in each stage. Otherwise, the comparisons will become unfair, that is, in the present constellation, T has a greater chance than R to be significantly superior to C.

The combining test statistic from (10), applied to the testing problem (77) above, takes on the value $Z_2^{RC}(0) = 2.16 + 1.77 = 3.93 > 3.237$, such that in a third step, H_0^{RC} could have been rejected at the second stage, too. Otherwise, the study had to be continued, provided we had included (77) in advance.

But in the present study, the interest is directed mainly towards the new drug T. The reference drug R is already on the market and had shown its superiority, when compared to placebo groups earlier in large studies. So also for safety concerns about the treatment, the number of observations on the new drug was chosen larger than on the reference drug. Since the treatment difference between T and C was expected to be larger than between T and $R - \Delta_0$, the control group C was chosen smaller than the reference group R but in a minimum relation to the test group T. So external considerations were more important than an optimal allocation of the sample sizes according to the formulae (72) to (75).

In the further analysis of the present example, the treatment effects are illustrated by the confidence intervals from (31). We obtain in the realized two stages of the study the following (≥ 0.95)-confidence intervals on the treatment difference $\mu_T - \mu_C$,

$$CI_1(\mu_T - \mu_C) = [0.10, 0.94] \quad \text{and} \quad CI_2(\mu_T - \mu_C) = [0.23, \ 0.83]$$

and (≥ 0.95)-confidence intervals on the treatment difference $\mu_T - \mu_R$ as

$$CI_1(\mu_t - \mu_R) = [-0.23, 0.41] \quad \text{and} \quad CI_2(\mu_T - \mu_R) = [-0.10, \ 0.36].$$

The simultaneous confidence level is at least 90% by Bonferroni's inequality.

Confidence intervals on the single parameters are provided by Section 3.5. We confine ourselves to the common variance parameter σ^2. Using the same critical values as above and the data from Table 1, we obtain (≥ 0.95)-confidence intervals on σ^2 by the approximation (51) as:

$$VI_1(\sigma^2)_{\text{appr}} = [0.781^2, 1.018^2] \quad \text{and} \quad VI_2(\sigma^2)_{\text{appr}} = [0.771^2, 0.939^2],$$

that is,

$$VCI_2(\sigma^2)_{\text{appr}} = VI_1(\sigma^2)_{\text{appr}} \cap VI_2(\sigma^2)_{\text{appr}} = [0.781^2, 0.939^2],$$

and using (47), by equating

$$V_1(\sigma^2) = \Phi^{-1}\left[F_{\chi^2_{200}}\left(200\ \frac{0.87^2}{\sigma^2}\right)\right] = \pm 2.289 \quad \text{and}$$

$$V_2(\sigma^2) = V_1(\sigma^2) + \Phi^{-1}\left[F_{\chi^2_{165}}\left(165\ \frac{0.81^2}{\sigma^2}\right)\right] = \pm 3.237,$$

and solving for σ^2, we compute the exact (≥ 0.95)-confidence intervals on σ^2 as

$$VI_1(\sigma^2) = [0.780^2, 0.982^2] \quad \text{and} \quad VI_2(\sigma^2) = [0.776^2, 0.920^2],$$

so that

$$VCI_2(\sigma^2) = VI_1(\sigma^2) \cap VI_2(\sigma^2) = [0.780^2, 0.920^2].$$

In the same way, let us consider only the parameter σ^2 for point estimation discussed in Section 4. The *approximate* median unbiased ML-estimates implied by (58) are

$$\widehat{\sigma^2}^{(2)}(1) = 0.87^2 \quad \text{and} \quad \widehat{\sigma^2}^{(2)}(2) = 0.8466^2$$

and by (59), we get the *meta-analytical* estimates

$$\widehat{\sigma^2}^{(3)}(1) = 0.87^2 \quad \text{and} \quad \widehat{\sigma^2}^{(3)}(2) = 0.8434^2.$$

The exact *median unbiased ML-estimates*, see (57), are obtained by equating $V_1(\sigma^2) = 0$ and $V_2(\sigma^2) = 0$ and solving for σ^2 as:

$$\widehat{\sigma^2}^{(1)}(1) = 0.8715^2 \quad \text{and} \quad \widehat{\sigma^1}^{(1)}(2) = 0.8428^2.$$

It should be noted, that by the early stopping of the study, we passed up the chance to improve the point and interval estimates in the third stage.

Further, we would like to remark, that the placebo arm C was not dropped after the first stage for several reasons. For example, the homogeneity, see (20), and eventual unwanted adverse side effects should be controlled in the following stages by a placebo group, too. Ethical and legal problems could be excluded.

Finally, it might be noted, that in the present study we had reliable a priori information, which, however, could be recognized earliest in the first interim analysis. So, when

we had conducted a non-adaptive group sequential trial by taking in the second stage the same sample sizes as computed by our formulae for the initial stage, we would had observed nearly surely quite similar results. Hence, the benefit of the adaptive design is here just having saved the costs for observing the difference of 35 patients.

But with non-reliable a priori information, the consequences might become quite different, see, for instance Hartung (2006), who points out ethical aspects and possible legal complications with non-adaptive designs, when, for instance, the treatment concerns a severe disease where patients cannot get back their status from baseline.

7 Final Remarks

In Section 5, we have computed sample sizes n using a normal approximation for applying t-variates. Nearly exact values are achieved by correcting the sample size n with the variance of a t_{n-1}-variate, that is, replacing n by $n_{\text{corr}} = n(n-1)/(n-3),\ n \geq 4$. The idea behind the correction is the same as in replacing a t-variate by a normal variate with identical variance. However, computed values have usually to be modified to fit some side conditions.

In Section 2.1 we have defined positive one-sided critical values $c\nu_j,\ j = 1, \ldots, K$, by the probability condition (11). For a fixed number of stages K and an overall significance level α, we get an O'Brien and Fleming (1979) design with constant critical values in (11), say $cv_j = cons_{OBF}(K, \alpha)$, and a Pocock (1977) design with monotone increasing critical values given as $cv_j = \sqrt{j}\ cons_{PO}(K, \alpha), j = 1, \ldots, K$, see Hartung (2006), where also some of these one-sided critical values are tabulated. Designs with intermediate values of the critical values are considered, for instance, in Jennison and Turnbull (2000).

Usually, two-sided critical values at level 2α for the corresponding symmetric two-sided tests are tabulated in literature. For $K \geq 2$, these two-sided critical values are slightly smaller than the one-sided critical values at level α. At least for $\alpha \leq 0.05$, these two-sided critical values may be used here for practical applications, see Jennison and Turnbull (2000, p. 192).

We have defined the two-sided confidence interval CI_k, see (31), as the intersection of the one-sided intervals $CI_{k,I}$ and $CI_{k,II}$, see (25) and (28), and the confidence coefficient of CI_k is at least $1 - 2\alpha$. If we use the critical values of the correspondent two-sided tests at level 2α, we 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 CI_K^0 reaches a confidence coefficient of exactly $1 - 2\alpha$.

However, using the lower boundary of CI_k^0 in the test decisions (27), the test level α cannot be guaranteed. Indeed, no severe differences are expected for practical applications at least for $\alpha \leq 0.05$, see above.

Moreover, let us consider the testing situation in a group sequential trial. In a superiority test, for example, the null-hypothesis is rejected at level α, if we observe $Z_{k^*}(0) > c\nu_{k^*}$ in at least one stage $k^* \in \{1, \ldots, K\}$, see (12). Usually the study is stopped after stage k^* because of having shown significance, see, for instance, Jennison and Turnbull (2000), Hartung (2006). Such a stop is a correct decision, since the study result cannot be reversed later in the same study. Consequently, when we continue the study, we have no risk to lose the already shown significance.

Suppose $k^* < K$ and the study is continued to reach a larger data base, for instance, for safety reasons in clinical trials, then we may observe $Z_k(0) \leq c\nu_k$ in all further stages $k > k^*$ without contradicting the already shown superiority. This fact is able to induce misunderstandings in practical applications caused by a lack of knowledge on the theoretical background. The same problem may arise, when, after shown significant noninferiority, the trial is continued for an attempt to reach superiority. Such possible misunderstandings are avoided by using CI_k, that means, its lower boundary, and the testing procedure (27). The automatically implied homogeneity test (20) by computing CI_k would react when quite different results would have been observed in later stages.

References

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

Brannath, W., Posch, M., Bauer, P. (2002). Recursive combination tests. *JASA* 97:236–244.

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

EMEA (The European Agency for the Evaluation of Medicinal Products) (1998). ICH Topic E9: Statistical Principles for Clinical Trials, London, CPMP/ICH/363/96.

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.

Hartung, J. (2006). Flexible Designs by adaptive plans of generalized Pocock- and O'Brien-Fleming-type and by Self-designing clinical trials. *Biometrical J.* 48:521–536.

Hartung, J., Böckenhoff, A., Knapp, G. (2003). Generalized Cochran-Wald statistics in combining of experiments. *J. Statist. Plann. Inferences* 113:215–237.

Hartung, J., Knapp, G. (2003). Confidence regions on variance components in an extended ANOVA model for combining information. *Acta Applicandae Mathematicae* 78:207–221.

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

Jennison, C. and Turnbull, B. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton and London: Chapman and Hall/CRC.

Koch, A. (2006). Confirmatory clinical trials with an adaptive design. *Biometrical J.* 48:574–585.

Maurer, W., Hothorn, A., Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: A-priori ordered hypotheses. In: Vollmer, J. (ed.) (1995). *Testing Principles in Clinical and Preclinical Trials*. New York: Gustav Fischer.

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

Pigeot, T., Schäfer, J., Röhmel, J., Hauschke, D. (2003). Assessing non-inferiority of a new drug in a three-arm clinical trial including a placebo. *Statist. Med.* 22:883–399.

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