Standardized Mean Differences in Adaptive Group Sequential Trials

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

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

Abstract: For studies comparing two independent groups, experimental E and control C, with normally distributed response variables, the outcome measure standardized differences of means is considered that is scale and translation invariant. This effect measure enables a convenient specification of a noninferiority margin in a concrete application. The present paper provides in particular a group sequential confidence interval approach to noninferiority trials and to switching between noninferiority and superiority for the effect size measure standardized mean difference. During the course of the trial, the sample size can be calculated in a completely adaptive way, based on the unblinded data of previously performed stages. Concrete rules for sample size updating are provided in this paper. Moreover, in each interim analysis, it is possible to change the planning from showing noninferiority to showing superiority or vice versa. A real data example is worked out in detail and the change in the planning from showing noninferiority to showing superiority is considered during the ongoing trial.

Keywords: Standardized difference of means, Effect size, Multi-stage confidence intervals, Adaptive sample size planning, Switching between noninferiority and superiority.

1 Introduction

In this paper, we consider comparative studies with normally distributed response variables in two independent groups, experimental E and control C. Common outcome measures are the difference of means, say $\mu_E - \mu_C$, the ratio of means, say μ_E/μ_C, and the standardized difference of means, say $(\mu_E - \mu_C)/\sigma$, where $\sigma^2 > 0$ denotes the common variance of the responses. In the analysis, the confidence interval approach is of particular attractiveness. It demonstrates in the best way a switching from noninferiority to superiority, if possible.

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

In a clinical examination, for example, when a new treatment is to be compared to a standard treatment with regard to noninferiority, the difficulty arises to choose not only the suitable outcome measure in advance but also the noninferiority margin before the beginning of the study. The noninferiority margin depends directly on the outcome measure. Since the difference of means is not scale invariant and the ratio of means is not translation invariant, the most convenient measure with respect to specifying a noninferiority margin in advance is the standardized difference of means that is both: scale invariant and translation invariant. That means, if both outcome variables, say X_E and X_C, are transformed by $aX_E + b$ and $aX_C + b$, $a \neq 1,\ b \neq 0$, which does not change anything from the clinical point of view, then only the outcome measure standardized difference of means remains unchanged.

Since the estimator of the standardized mean difference follows a scaled noncentral t-distribution, it is not widely used in clinical trials although it would be the suitable outcome measure in many cases. An exact confidence interval for the effect size standardized mean difference is discussed by Hedges and Olkin (1985). In spite of its practical importance, the standardized mean difference does not seem to be considered in group sequential trials until now, neither for testing noninferiority nor for deriving confidence intervals.

In this paper, we consider the outcome measure standardized mean difference in general adaptive group sequential trials, see Hartung (2006). Parameterized p-values, see Cox and Hinkley (1974), of the several stages are combined by the inverse normal method from meta-analysis, see Hedges and Olkin (1985), Hartung, Knapp, and Sinha (2008). As with the confidence interval of Hedges and Olkin, the proposed confidence intervals are defined implicitly and for obtaining the boundaries, nonlinear equations have to be solved. Indeed, the solutions are always unique. Besides this, we provide approximate confidence intervals in an explicit form.

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

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

The outline of the present paper is as follows: In Section 2.1, one-sided group sequential confidence intervals for the standardized mean difference are derived, and switching between noninferiority and superiority is considered. In Section 2.2, two-sided confidence intervals and a test on the homogeneity of the standardized mean differences underlying the different stages of the trial are presented. In Section 2.3, approximate confidence intervals are derived in an explicit form. Section 3 contains median unbiased maximum likelihood estimators for the standardized mean difference at each stage. Section 4 deals with general adaptive sample size planning. Section 5 contains a real data example in an adaptive three-stage Pocock (1977) design, which is worked out in detail and demonstrates switching from noninferiority to superiority during the ongoing trial. Some additional comments are given in Section 6.

2 Nested Multi-Stage Confidence Intervals for the Standardized Difference of Normal Means

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 $\sigma^2 > 0$, succinctly, $X_E \sim \mathcal{N}(\mu_E, \sigma^2)$ and $X_C \sim \mathcal{N}(\mu_C, \sigma^2)$.

Let $\Delta_0 \geq 0$ be a margin for the noninferiority parameter $\Delta \geq 0$. We are interested in hypotheses testing for noninferiority when the noninferiority margin for $\mu_E - \mu_C$ is put in relation to the standard deviation, that is, the test problem is

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

at a predefined level α, $0 < \alpha < 1/2$. The alternative stands for ($\Delta\sigma$-)noninferiority, $0 < \Delta \leq \Delta_0$, and means superiority of the experimental group E with regard to the control group C for $\Delta = 0$. Let ϑ denote the standardized difference of the means, say

$$\vartheta = \frac{\mu_E - \mu_C}{\sigma},$$

we can reformulate the hypotheses in (1) as follows,

$$\mathrm{H}_{0,\Delta}: \ \vartheta + \Delta = 0 \quad \text{versus} \quad \mathrm{H}_{0,\Delta}: \ \vartheta + \Delta > 0, \quad 0 \leq \Delta \leq \Delta_0. \tag{2}$$

We consider a comparative study which is carried out consecutively in a number, say K, of independent stages. In the i-th stage, $i = 1, \ldots, K$, we observe the sample mean $\bar{X}_{E_i}$ of $n_{E_i} \geq 2$ responses, the sample mean $\bar{X}_{C_i}$ of $n_{C_i} \geq 2$ responses, and the pooled sample variance S_i^2 in the two independent groups E and C. Note that S_i^2 is stochastically independent of the sample means and follows a scaled χ^2-distribution with $\nu_i = n_{E_i} + n_{C_i} - 2$ degrees of freedom, that is,

$$\nu_i \, \frac{S_i^2}{\sigma^2} \sim \chi^2_{\nu_i}, \quad \nu_i = n_{E_i} + n_{C_i} - 2, \quad i = 1, \ldots, K. \tag{3}$$

2.1 Nested One-sided Confidence Intervals

The standardized mean difference ϑ is estimated by use of Hedges's g which is given by

$$g_i = \frac{\bar{X}_{E_i} - \bar{X}_{C_i}}{S_i} \tag{4}$$

in the i-th stage, $i = 1, \ldots, K$. The estimator g_i from (4) possesses the distributional property that

$$\sqrt{b_i}\; g_i \sim t\left(\nu_i, \sqrt{b_i}\; \vartheta_i\right), \quad b_i = \frac{n_{E_i}\, n_{C_i}}{n_{E_i} + n_{C_i}}, \quad \nu_i = n_{E_i} + n_{C_i} - 2, \quad i = 1, \ldots, K, \tag{5}$$

where $t(\nu_i, \ \sqrt{b_i}\; \vartheta)$ stands for the noncentral t-distribution with ν_i degrees of freedom and noncentrality parameter $\sqrt{b_i}\; \vartheta$, see Hedges (1981), Hedges and Olkin (1985).

Let $F_{t(\nu_i, \ \sqrt{b_i}\ \vartheta)}$ denote the cumulative distribution function of a $t(\nu_i, \ \sqrt{b_i}\; \vartheta)$-variate, then, with the *true* parameter ϑ, it holds for the $1-p$-value

$$F_{t(\nu_i, \ \sqrt{b_i}\ \vartheta)}\left(\sqrt{b_i}\; g_i\right) \sim \mathcal{U}(0,1), \quad i = 1, \ldots, K, \tag{6}$$

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

$$z_i(\vartheta) := \Phi^{-1}\left[F_{t(\nu_i, \ \sqrt{b_i}\ \vartheta)}(\sqrt{b_i}\; g_i)\right] \sim \mathcal{N}(0,1), \quad i = 1, \ldots, K, \tag{7}$$

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

Since the stages of the study are assumed to be independent, we can define up to the j-th stage the combining pivotal statistic as

$$Z_j(\vartheta) := \sum_{i=1}^{j} z_i(\vartheta) \sim \sqrt{j}\ \mathcal{N}(0,1), \quad j = 1, \ldots, K. \tag{8}$$

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

$$\mathrm{P}\left(\sum_{i=1}^{j} Y_i \leq cv_j \text{ for all } j = 1, \ldots, K\right) = 1 - \alpha, \tag{9}$$

see Hartung (2006). Using critical values cv_j defined by (9), we get the following probability statements for the combining pivotal statistics from (8),

$$\mathrm{P}_\vartheta\left(Z_j(\vartheta) \leq cv_j \text{ for } j = 1, \ldots, k \leq K\right) \begin{cases} \geq 1 - \alpha & \text{for } k < K, \\ = 1 - \alpha & \text{for } k = K. \end{cases} \tag{10}$$

From (10), we define the lower confidence sets on ϑ as

$$CI_{k,L}(\vartheta) := \{\tilde{\vartheta} \in I\!R \mid Z_j(\tilde{\vartheta}) \leq cv_j \text{ for } j = 1, \ldots, k\}, \quad k = 1, \ldots, K. \tag{11}$$

The confidence sets in (11) are nested, that is, $CI_{k+1,L}(\vartheta) \subset CI_k(\vartheta)$, $k = 1, \ldots, K-1$, and, by (10), the confidence coefficient of $CI_{k,L}(\vartheta)$ is at least $1-\alpha$, and exactly $1-\alpha$ for $k = K$.

The distribution function of the noncentral t-distribution in (strictly) monotone decreasing with respect to the noncentrality parameter, that is, $\vartheta_1 > \vartheta_2$ implies

$$F_{t(\nu_i,\ \sqrt{b_i}\vartheta_1)}(y) < F_{t(\nu_i,\ \sqrt{b_i}\vartheta_2)}(y) \quad \forall\ y \in I\!R.$$

Further, $\Phi^{-1}(.)$ is a monotone increasing function in its argument, implying that $z_i(\vartheta)$ from (7) is monotone decreasing in ϑ. Consequently, the combining statistics $Z_j(\vartheta)$, $j = 1, \ldots, K$, from (8) are monotone decreasing in ϑ. Thus $CI_{k,L}(\vartheta)$ can be represented as a genuine interval, that is,

$$CI_{k,L}(\vartheta) = [\vartheta_{Lk}, \infty) \tag{12}$$

where $\vartheta_{Lk} = \max\{\vartheta_L(1), \ldots, \vartheta_L(k)\}$ and $\vartheta_L(j)$ solves

$$Z_j(\vartheta_L(j)) = cv_j, \quad j = 1, \ldots, k,\ k = 1, \ldots, K. \tag{13}$$

Note, that the solutions $\vartheta_L(j)$ in (13) are unique and can be iteratively found, for instance, by the bisection method.

Let us apply the multi-stage confidence intervals from (12) to the testing problem (1) at level of at most α. At stage k, $k = 1, \ldots, K$, using ϑ_{Lk} from (12), we proceed as follows:

$$\begin{aligned} &\text{if } -\Delta < \vartheta_{Lk}, \text{ then we decide for } \mathrm{H}_{1,\Delta}, \\ &\text{if } -\Delta_0 \geq \vartheta_{Lk}, \text{ then stay with } \mathrm{H}_{0,\Delta_0} \end{aligned} \tag{14}$$

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

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

2.2 Nested Two-sided Confidence Intervals and Homogeneity of Effect Sizes

In analogy to (11), let us define the upper confidence sets on ϑ as

$$CI_{k,U}(\vartheta) := \{\tilde{\vartheta} \in I\!R \mid -cv_j \leq Z_j(\tilde{\vartheta}) \text{ for } j = 1, \ldots, k\}, \quad k = 1, \ldots, K. \tag{15}$$

Again, the confidence sets are nested, that is, $CI_{k+1,U}(\vartheta) \subset CI_{k,U}(\vartheta)$, $k = 1, \ldots, K-1$, and each confidence set has a confidence coefficient of at least $1-\alpha$, being exactly $1-\alpha$ for $k = K$. The interval representation is given by

$$CI_{k,U}(\vartheta) = (-\infty, \vartheta_{Uk}], \tag{16}$$

where $\vartheta_{Uk} = \min\{\vartheta_U(1), \ldots, \vartheta_U(k)\}$ and $\vartheta_U(j)$ solves uniquely

$$Z_j(\vartheta_U(j)) = -cv_j, \quad j = 1, \ldots, k, \; k = 1, \ldots, K. \tag{17}$$

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

$$CI_k(\vartheta) := CI_{k,L}(\vartheta) \cap CI_{k,U}(\vartheta) = [\vartheta_{Lk}, \vartheta_{Uk}] \tag{18}$$

with ϑ_{Lk} from (12) and ϑ_{Uk} from (16). The confidence intervals are nested, that is,

$$CI_{k+1}(\vartheta) \subset CI_k(\vartheta), \quad k = 1, \ldots, K-1, \tag{19}$$

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

Denote $I_k(\vartheta) = [\vartheta_L(k), \vartheta_U(k)]$, see (13) and (17), the individual confidence interval on ϑ at the k-th stage. Then it holds

$$CI_1(\vartheta) = I_1(\vartheta) \quad \text{and} \quad CI_k(\vartheta) = CI_{k-1}(\vartheta) \cap I_k(\vartheta), \quad k = 2, \ldots, K. \tag{20}$$

Since $CI_k \subset I_k$, the interval $I_k(\vartheta)$ is another two-sided confidence interval with confidence coefficient of at least $1-2\alpha$ on ϑ. The interval $I_k(\vartheta)$ results from the boundaries in stage k alone and will be always nonempty. Therefore, $I_k(\vartheta)$ may be preferred to $CI_k(\vartheta)$, see for instance Jennison and Turnbull (2000, p. 192) in their corresponding settings.

Depending on the choice of α, the two-sided confidence interval CI_k from (18) can be empty, that is, it may occur that $\vartheta_{Uk} < \vartheta_{Lk}$. For interpreting such an event, let us consider the extended model that each stage of the study has an individual parameter, say $\vartheta_i = (\mu_{E_i} - \mu_{C_i})/\sigma_i,\ i = 1, \ldots, K$. Since Hedges's g_i from (4) estimates ϑ_i, we have $\sqrt{b_i}\ g_i \sim t(\nu_i,\ \sqrt{b_i}\ \vartheta_i)$, see (5). In analogy to (7), we get $z_i(\vartheta_i) \sim \mathcal{N}(0,1),\ i = 1, \ldots, K$, and

$$Z_j(\vartheta_1, \ldots, \vartheta_j) := \sum_{i=1}^{j} z_i(\vartheta_i) \sim \sqrt{j}\ \mathcal{N}(0,1),\ j = 1, \ldots, K, \tag{21}$$

for the combining pivotal statistic up to the j-th stage, see (8).

Denote $d' = (d_1, \ldots, d_k)$ the transposed of a vector d in $I\!R^k$, then, by (9), the k-dimensional confidence region, $k = 1, \ldots K$,

$$CR_k := \{d \in I\!R^k | -cv_j \le Z_j(d_1, \ldots, d_j) \le cv_j \quad \text{for } j = 1, \ldots, k\} \tag{22}$$

covers $(\vartheta_1, \ldots, \vartheta_k)'$ with probability of at least $1-2\alpha,\ 0 < \alpha < 1/2$. Note that CR_k is not empty for all $\alpha \in (0, 1/2)$. For example, the realized vector of $\hat{\vartheta}_i$ defined by $z_i(\hat{\vartheta}_i) = 0$, $i = 1, \ldots, k$, lies always in CR_k, where $\hat{\vartheta}_i$ is the median unbiased maximum likelihood estimator of ϑ_i in the i-th stage, see Section 3.

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

In testing

$$\mathrm{H}_{0,\mathrm{hom}}(k): \ \vartheta_1 = \ldots = \vartheta_k \quad \text{versus} \quad \mathrm{H}_{1,\mathrm{hom}}(k): \ \vartheta_{i_1} \neq \vartheta_{i_2} \tag{23}$$

for some $i_1, \ i_2 \in \{1, \ldots, k\}, \ k = 2, \ldots, K$, the homogeneity hypothesis $\mathrm{H}_{0,\mathrm{hom}}(k)$ will be rejected at level of at most 2α if the two-sided confidence interval $CI_k(\vartheta)$ from (18) is empty. If $\mathrm{H}_{0,\mathrm{hom}}(k^*)$ is rejected, then also $\mathrm{H}_{0,\mathrm{hom}}(k)$ for $k^* \leq k \leq K$. An alternative to this homogeneity test does not seem to be known.

Suppose that up to stage $k-1$ the intersections in (20) are nonempty and in the k-th stage, CI_k is empty for a common level α, that is, the nonempty interval I_k lies completely outside the nonempty interval CI_{k-1}. Therefore, up to an error rate of 2α, see (23), we may consider that a break in the underlying standardized mean differences has been observed. So, with regard to statistical concerns, results from this stage k should not influence conclusions from the previous stages. Consequently, preferring I_k to CI_k does not provide any real advantage. On the other hand, under the model assumption of an identical standardized mean difference underlying the different stages of the study, the probability to obtain an empty confidence interval CI_k is bounded by 2α.

Finally, we would like to remark that, in the case $K = 1$, the interval from (18) is the exact confidence interval discussed by Hedges and Olkin (1985, p. 91).

2.3 Approximative Confidence Intervals

From Hedges and Olkin (1985, Chapter 5) or Hartung, Knapp, and Sinha (2008, Chapter 2), we take over the following approximations. In the i-th stage, see (4),

$$g_i^* = \left(1 - \frac{3}{4n_i - 9}\right) g_i, \quad n_i = n_{E_i} + n_{C_i}, \quad i = 1, \ldots, K, \tag{24}$$

is an approximately unbiased estimator of the standardized mean difference ϑ. The variance of g_i, or g_i^*, is approximately unbiasedly estimated in the i-th stage by, see (5),

$$V_i = \frac{1}{b_i} + \frac{g_i^2}{2\nu_i}, \quad \frac{1}{b_i} = \frac{1}{n_{E_i}} + \frac{1}{n_{C_i}}, \quad \nu_i = n_i - 2, \quad i = 1, \ldots, K, \tag{25}$$

and g_i is approximately normally distributed, so that $z_i(\vartheta)$ in (7) is approximated by

$$z_i^A(\vartheta) = \Phi^{-1}\left[\Phi\left((g_i^* - \vartheta)/\sqrt{V_i}\right)\right].$$

That means, the combining statistic $Z_j(\vartheta)$ in (8) is approximated by

$$Z_j^A(\vartheta) = \sum_{i=1}^{j} \frac{g_i^* - \vartheta}{\sqrt{V_i}} \quad \overset{\text{appr.}}{\sim} \quad \sqrt{j}\,\mathcal{N}(0,1), \quad j = 1, \ldots, K. \tag{26}$$

Equating $Z_j^A(\vartheta)$ to cv_j, see (13), and to $-cv_j$, see (17) and solving for ϑ, yields the approximate individual confidence interval at the j-th stage, see (20),

$$I_j^A(\vartheta): \quad \sum_{i=1}^{j} \frac{g_i^*/\sqrt{V_i}}{\sum_{h=1}^{j} 1/\sqrt{V_h}} \pm \frac{cv_j}{\sum_{h=1}^{j} 1/\sqrt{V_h}}, \quad j = 1, \ldots, K. \tag{27}$$

By setting $CI_1^A = I_1^A$ and $CI_k^A = CI_{k-1}^A \cap I_k^A$, $k = 2, \ldots, K$, we get approximations of the confidence intervals CI_k in (18). The boundaries of these approximative confidence intervals may be used as starting values in an iterative procedure to determine the exact confidence intervals.

3 Point Estimation of the Standardized Mean Difference

The combining statistic $Z_j(\vartheta)$ from (8) is $\mathcal{N}(0, j)$-distributed with mode and median 0. So, the maximum likelihood (ML) estimator $\hat{\vartheta}_{\mathrm{ML}}(j)$ of the standardized mean difference ϑ at stage j is given by:

$$\hat{\vartheta}_{\mathrm{ML}}(j) \text{ solves } Z_j\left(\hat{\vartheta}_{\mathrm{ML}}(j)\right) = 0, \quad j = 1, \ldots, K, \tag{28}$$

where the solution in (28) is unique.

The *global* p-value at stage j is

$$p_G(j) = 1 - \Phi\left[Z_j(\vartheta)/\sqrt{j}\right], \quad j = 1, \ldots, K, \tag{29}$$

and solving (29) for $p_G(j) = 1/2$ yields $\hat{\vartheta}_{\mathrm{ML}}(j)$ as solution. Note, that $Z_j(\vartheta)$ is monotone in ϑ. Consequently, see Cox and Hinkley (1974, p. 273),

$$\hat{\vartheta}_{\mathrm{ML}}(j) \quad \text{is median unbiased.} \tag{30}$$

That means, $\hat{\vartheta}_{\mathrm{ML}}(j)$ lies with equal probability as well below the parameter ϑ as above ϑ.

Using in (28) the approximate combining statistic $Z_j^A(\vartheta)$ from (26), we get the approximate median unbiased ML-estimator at stage j as, see (24) and (25),

$$\hat{\vartheta}_{\mathrm{ML}}^A(j) = \sum_{i=1}^{j} \frac{g_i^*/\sqrt{V_i}}{\sum_{h=1}^{j} 1/\sqrt{V_h}}, \quad j = 1, \ldots, K. \tag{31}$$

Note that the stage based estimators of ϑ are weighted by the inverses of their estimated standard errors in (31) and not by the inverses of their estimated variances as known from meta-analysis, see for instance Hartung, Knapp, and Sinha (2008).

The standard meta-analytical estimator up to the j-th stage takes on the form

$$\hat{\vartheta}_{\mathrm{MA}}(j) = \sum_{i=1}^{j} \frac{g_i^*/V_i}{\sum_{h=1}^{j} 1/V_h}, \quad j = 1, \ldots, K. \tag{32}$$

When sample sizes are chosen adaptively and the end of the study depends on a testing decision, no special approximate properties of this estimator are known so far, in contrary to the estimator $\hat{\vartheta}_{\mathrm{ML}}(j)$ from (28) or its approximation $\hat{\vartheta}_{\mathrm{ML}}^A(j)$. Weighted means like $\hat{\vartheta}_{\mathrm{ML}}^A(j)$ from (31) are used in the generalized Cochran-Wald statistics considered by Hartung, Böckenhoff, and Knapp (2003).

4 Adaptive Sample Size Planning for the Standardized Mean Difference

Planning with equal sample sizes in the two groups and suppressing the subscript i, we set $n_E = n_C = m$ and get from (25) for the approximate variance of g,

$$V_0 = \frac{1}{m}\left(2 + \frac{\vartheta^2}{4 - 4/m}\right) =: \frac{1}{m}\, v(m). \tag{33}$$

With some initial fixed m_0, let us define the random variables

$$X \sim \mathcal{N}\left(\vartheta, v(m_0)\right) \quad \text{and} \quad \bar{X} \sim \mathcal{N}\left(\vartheta, \frac{1}{m} v(m_0)\right). \tag{34}$$

For fixed $\Delta \in [0, \Delta_0]$ and ϑ^* with $\vartheta^* + \Delta > 0$, we want to test the point hypotheses $\mathrm{H}_0^*: \ \vartheta + \Delta = 0$ versus $\mathrm{H}_1^*: \ \vartheta + \Delta = \vartheta^* + \Delta > 0$ by use of the test statistic

$$T = \sqrt{m}\frac{\bar{X} + \Delta}{\sqrt{v(m_0)}} \sim \mathcal{N}(0,1) \text{ under } \mathrm{H}_0^*. \tag{35}$$

Assume the test is carried out in the j-th stage and $\vartheta^* = \hat{\vartheta}(j-1)$, where $\hat{\vartheta}(j-1)$ is some estimate of ϑ, see Section 3, based on the information of stage 0, stage 1, …, stage $j-1$, satisfying $\hat{\vartheta}(j-1) + \Delta > 0$. Hereby, stage 0 stands for a priori information.

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

$$m = f_{j-1}(\alpha, \beta, \Delta) := \frac{[\max\{0, \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}]^2}{\left(\hat{\vartheta}(j-1) + \Delta\right)^2 \ / \ [2 + \hat{\vartheta}(j-1)^2/(4 - 4/m_0)]}, \tag{36}$$

with $\hat{\vartheta}(j-1) + \Delta > 0$, $j = 1, \ldots, K$. Note that, for ease of presentation, we use a normal sample size spending function in (36). Furthermore, we may replace m_0 by m in (36) and iterate until we reach a stabilization in the sense that two following values differ less than 1. The statistic T in (35) corresponds to the approximate test statistic in the i-th stage for $m = n_{E_i} = n_{C_i}$, see (4) and (25),

$$T_i = \frac{g_i + \Delta}{\sqrt{1/b_i + g_i^2/2\nu_i}}, \quad i = 1, \ldots, K, \tag{37}$$

which is approximately $\mathcal{N}(0,1)$-distributed under H_0^*. Indeed, T_i is used here only for deriving the above formula (36).

Recall now from (9) that it holds:

$$\begin{aligned}\left\{\sum_{i=1}^{h} Y_i \le cv_h \text{ for } h = 1, \ldots, j-1 \text{ and } \sum_{i=1}^{j-1} Y_i + \sqrt{K-(j-1)}Y_j \le cv_K\right\} \\ \supset \left\{\sum_{i=1}^{h} Y_i \le cv_h \text{ for all } h = 1, \ldots, K\right\}.\end{aligned} \tag{38}$$

In the group sequential trial, the hypothesis $\mathrm{H}_{0,\Delta}$, see (2), will be rejected if $Z_j(-\Delta) > cv_j$, see (8) and (10). Then, by (38), if we decide after stage $j-1$ to omit the interim analyses j up to $K-1$, we can assign the remaining weight $\sqrt{K-(j-1)}$ to the next final study part and build the final test statistic, see (7) and (8), as

$$Z_{j,K}(-\Delta) := Z_{j-1}(-\Delta) + \sqrt{K-(j-1)}\ \Phi^{-1}\left[F_{t\left(\nu_j, \sqrt{b_j}(-\Delta)\right)}\left(\sqrt{b_j}g_j\right)\right], \tag{39}$$

where $Z_{j,K}(-\Delta) \sim \sqrt{K}\ \mathcal{N}(0,1)$ under $\mathrm{H}_{0,\Delta}$, $j = 1, \ldots, K$, defining $Z_0 = 0$. The test statistic $Z_{j,K}(-\Delta)$ has to be compared with the K-th critical value cv_K in testing $\mathrm{H}_{0,\Delta}$.

Note, that the p-value of testing $\mathrm{H}_{0,\Delta}$ at stage i by use of $\sqrt{b_i}g_i$ is given by, see (5) and (6),

$$p_i = p_i(\Delta) = 1 - F_{t\left(\nu_i, \sqrt{b_i}(-\Delta)\right)}\left(\sqrt{b_i}g_i\right), \quad i = 1, \ldots, K. \tag{40}$$

Assume that a significant result has not been obtained up to stage $j-1$, that is, $Z_i(-\Delta) \leq cv_i$ for $i = 1, \dots, j-1$. In the next study part we want to reach cv_K by use of the final test statistic

$$\hat{Z}_{j,K}(-\Delta) := Z_{j-1}(-\Delta) + \sqrt{K-(j-1)}\ \Phi^{-1}[1 - \hat{p}_{j,K}(\Delta)], \tag{41}$$

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

$$\hat{p}_{j,K}(\Delta) = 1 - \Phi[(cv_K - Z_{j-1}(-\Delta))/\sqrt{K-(j-1)}]. \tag{42}$$

To detect a deviation of the null-hypothesis in the direction $\mathrm{H}_{1,\Delta}$ at $\vartheta + \Delta = \hat{\vartheta}(j-1) + \Delta > 0$ with the (conditional) power $1-\beta$, the sample size for the next final study part must be chosen with (36) in each group as

$$M_j = M_j(\Delta) = f_{j-1}(\hat{p}_{j,K}(\Delta), \beta, \Delta), \quad j = 1, \dots, K. \tag{43}$$

Consequently, $\hat{p}_{j,K}$ can be named as a conditional error function.

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 choose the sample size in each group for stage j proportionally as

$$n_j/2 = n_j(\Delta)/2 = \frac{M_j(\Delta)}{K-j+1}, \quad n_{E_j} = n_{C_j} \approx n_j/2, \quad j = 1, \dots, K, \tag{44}$$

which is a (slightly) conservative choice by (38), and use cv_j as critical value for $Z_j(\vartheta)$ in stage j. Note that each sample size should be at least 2.

Especially for $j = 1$, we get the starting sample size of the trial as

$$n_1 = \frac{2M_1}{K},\ M_1 = f_0(\hat{p}_{1,K}, \beta, \Delta), \quad \hat{p}_{1,K} = 1 - \Phi(cv_K/\sqrt{K}), \tag{45}$$

where in f_0 from (36), we use some prior guess $g_0 = \hat{\vartheta}(0)$ of ϑ with $g_0 + \Delta > 0$.

We start with n_1 observations in the first stage, n_1 from (45). Then with the proceeding above, we reach the full power $1-\beta$, conditioned on $\vartheta + \Delta = \hat{\vartheta}(K-1) + \Delta > 0$, latest in stage $j = K$, if not stopped before because of shown significance,. Note, that the estimates $\hat{\vartheta}(j-1)$ are used only for planning the sample sizes, but not for computing the confidence intervals. When we replace $Z_j(\vartheta)$ by $Z_j^A(\vartheta)$ from (26) in the above considerations, we obtain an approximative proceeding.

Further, we may formally define the p-values, see (6), as suiting to the null-hypothesis that ϑ is the *true* parameter, see Cox and Hinkley (1974, p. 221). So, we may apply

Table 1: Controlled clinical trial concerning patients with acne papulopustulosa in an adaptive 3-stage Pocock (1977) design with early stop for superiority after stage 2 at given one-sided significance level $\alpha = 0.005$.

Stage	Adaptive sample size	Data on $\vartheta = (\mu_E - \mu_C)/\sigma$	ML-estimate	Confidence interval
i	n_i	g_i	$\hat{\vartheta}_{ML}(i)$	$CI_i(\vartheta)$
0	-	0.8	0.8	[Level $\geq$ 0.99]
1	24	1.177	1.1230	$[-0.1425, 2.3992]$
2	12	1.073	1.0572	$[\ \ 0.0136, 2.1076]$
3	STOP Because of shown superiority			

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 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).

5 A Real Data Example

Using the raw difference of means as effect measure in a controlled clinical trial concerning patients with acne papulopustulosa, Lehmacher and Wassmer (1999) discussed an adaptive three-stage group sequential test of Pocock (1977) type, which led to an early stop for superiority of the experimental group E with respect to the control group C after the second stage at the one-sided overall significance level of $\alpha = 0.005$. The response variable was the reduction of bacteria (after six weeks of treatment) from baseline, examined on agar plates and measured as $\log CFU/cm^2$ (CFU: colony forming units). We take over the parameter estimates and compute the observed standardized means g_i as presented in Table 1.

Assuming that the noninferiority boundary for the treatment difference $\mu_E - \mu_C$ is specified in relation to the standard deviation, say $\Delta_0\sigma$, with $\Delta_0 = 20\%$, we apply the methods proposed in the previous sections. We choose the same test level $\alpha = 0.005$ and the power $1-\beta = 0.90$. Each stage is planned with equal sample sizes in both groups, where the two involved drugs will be equally randomized within blocks of size 6.

For a three-stage Pocock design, we get the (one-sided) critical values $cv_j = 2.873\sqrt{j}$, $j = 1, 2, 3$, using the combining statistic (8) in (9) for $\alpha = 0.005$, see Hartung (2006, p. 533), or Jennison and Turnbull (2000, p. 26) for the two-sided level 0.01. Planning with $\Delta = 0.2$ for showing noninferiority, we compute the value $n_1 = 24.9$ for the total sample size of the first stage in (45) with the prior guess $g_0 = 0.8$ for the standardized mean difference ϑ, see Table 1. Starting with $m_0 = 30$ in (36) we only need one iteration. Because of the block size 6, we begin the trial with $n_{E_1} = n_{C_1} = 12$ patients.

For showing superiority we would calculate at least 19.3 patients in each group at the first stage. For comparison, in a one-stage trial, we would compute a fixed sample size in each group of at least 32.2 for showing noninferiority and of at least 50.3 for showing superiority. By $50.3 \times 1.15/3 = 19.3$, for instance, we confirm the rule of thumb that a Pocock design needs about 5% per stage additional subjects when compared to a non-group sequential trial, see Jennison and Turnbull (2000, p. 27).

Beginning with $n_1 = 24$ patients, we observe the estimate $g_1 = 1.177$ for the standardized mean difference, see Table 1. For a detailed illustration, let us demonstrate the approximate procedure. With $g_1^* = 1.13$, see (24), and $V_1 = 0.198$, see (25), we get the first approximate confidence interval on ϑ of size ≥ 0.99, see (27), as

$$I_1^A(\vartheta) = [-0.142 \quad , \quad 2.414] \,,$$

which lies clearly above $-\Delta_0 = -0.2$. At level of at most 0.005, the significant noninferiority of the experimental group E with regard to the control group C has been already shown, see (14).

In the further planning we switch to showing superiority, that is, we set $\Delta = 0$ in the following. Because the a priori information from Table 1 does not seem to be reliable, we use $g_1 = 1.177$ as $\hat{\vartheta}(1)$ in the sample size spending function f_1 from (36). With $Z_1^A(0) = 2.553$, see (26), we compute the projected p-value, see (42),

$$\widehat{p_{2,3}}(0) = 1 - \Phi\left[\left(2.873\sqrt{3} - 2.553\right)/\sqrt{2}\right] = 1 - \Phi[1.71345].$$

Then we obtain the value

$$2\ M_2(0) = 2\ \frac{[1.71345 + 1.28]^2}{1.177^2}\left(2 + \frac{1.177^2}{4 - 4/12}\right) = 30.76$$

by (43) and (36) for the total sample size of the remaining two stages, where the m_0 in (36) is set equal to 12. Instead of the projected 15 or 16 observations for the second stage, see (44), the decision was made for $n_2 = 12$ patients to be observed in the second stage because of the block size 6 and the option to carry out a third stage if the result of the second stage would not be satisfying. Note that by (14), the already shown noninferiority remains valid.

In the second stage, we observe the estimate $g_2 = 1.073$ for the standardized mean difference, see Table 1. With $g_2^* = 0.990$ and $V_2 = 0.391$, see (24), (25), $1/\sqrt{V_1} + 1/\sqrt{V_2} = 3.847$, $g_1^*/\sqrt{V_1} + g_2^*/\sqrt{V_2} = 4.136$, we obtain the second approximate confidence interval on ϑ of size ≥ 0.99, see (27), as

$$I_2^A(\vartheta) : \frac{4.136}{3.847} \pm \frac{2.873 \cdot \sqrt{2}}{3.847} = [0.019 \quad , \quad 2.131],$$

which lies above 0. By (14) based on this interval, the trial is stopped after the second stage because of having shown the superiority of the experimental group E with respect to the control group C at level of at most 0.005.

The midpoints of the above intervals are the approximate median unbiased ML-estimates for ϑ up to the corresponding stages, see (31),

$$\hat{\vartheta}_{ML}^A(1) = 1.136 \quad \text{and} \quad \hat{\vartheta}_{ML}^A(2) = 1.075.$$

With the exact combining statistics, see (7),

$$Z_1(\vartheta) = \Phi^{-1}\left[F_{t(22,\sqrt{6}\ \vartheta)}\left(1.177\sqrt{6}\right)\right], \text{ and}$$
$$Z_2(\vartheta) = Z_1(\vartheta) + \Phi^{-1}\left[F_{t(10,\sqrt{3}\ \vartheta)}\left(1.073\sqrt{3}\right)\right],$$

we obtain the exact confidence intervals for ϑ by equating $Z_1(\vartheta)$ to ± 2.873 and $Z_2(\vartheta)$ to $\pm 4.063 = \pm 2.873\sqrt{2}$ and solving for ϑ. Equating $Z_1(\vartheta) = 0$ and $Z_2(\vartheta) = 0$ and solving for ϑ yield the exact median unbiased ML-estimates $\hat{\vartheta}_{ML}(j), j = 1, 2$, for the standardized mean difference. The exact results are put together in Table 1.

6 Final Remarks

In Section 2.1, we have defined positive one-sided critical values cv_j, $j = 1, \ldots K$, by the probability condition (9). 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 (9), 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 correspondent 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 as the intersection of the one-sided intervals $CI_{k,L}$ and $CI_{k,U}$, see (18), 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, CI_K^0 reaches a confidence coefficient of exactly $1 - 2\alpha$. However, using the lower boundary of CI_k^0 in the test decision (14), 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 $H_{0,0}$ is rejected at level α in favor of $H_{1,0}$ if we observe $Z_{k^*}(0) > cv_{k^*}$ in at least one stage $k^* \in \{1, ..., K\}$ or that the individual confidence interval $I_{k^*}(\vartheta)$, see (20), lies above 0, as implied by (9). 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 $Z_k(0) \leq cv_k$ in all further stages $k > k^*$ or that I_k covers 0 without contradicting the already shown superiority. This fact is able to induce misunderstandings in practical applications caused by a lack of knowledge of the theoretical background. The same problem may arise when, after a shown significant noninferiority, the trial is continued for an attempt to reach superiority. Such possible misunderstandings are avoided by using CI_k instead of I_k as proposed in the testing procedure (14); see also the discussion on the use of I_k and CI_k in Section 2.2. The automatically implied homogeneity test (23) would react when quite different results would have been observed in later stages.

References

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

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

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

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 Journal* **48**, 521–536.

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.

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

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators, *Journal of Educational Statistics* **6**, 107–128.

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

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

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

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

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