Properties of a Confirmatory Two-Stage Adaptive Procedure for Assessing Average Bioequivalence

Marie Louise Østerdal\textsuperscript{a}, Kyle Raymond\textsuperscript{a}, and Christian Bressen Pipper\textsuperscript{b,c}

\textsuperscript{a}Department of Biostatistics, Leo Pharma A/S, Ballerup, Denmark; \textsuperscript{b}Biostatistics OSCD & Outcomes 1, Novo Nordisk A/S, Soeborg, Denmark; \textsuperscript{c}Department of Public Health, University of Southern Denmark, Odense, Denmark

\begin{abstract}
We investigate a confirmatory two-stage adaptive procedure for assessing average bioequivalence and provide some insights to its theoretical properties. Effectively, we perform Two One-Sided Tests (TOST) to reach an overall decision about each of the two traditional null-hypotheses involved in declaring average bioequivalence. The tests are performed as combination tests separately for each hypothesis based on the corresponding pair of stagewise \( p \)-values. Features of the procedure include a built in futility, sample size reassessment, and the ability to simultaneously assess average bioequivalence with respect to multiple endpoints while controlling the familywise error rate. To facilitate inference at the end of a trial we consider confidence limits that match the decision reached on each one sided hypothesis and provide theory ensuring their appropriateness. The performance is assessed by simulation in the context of planning a study to compare two different administrations of an antibody treatment of atopic dermatitis.
\end{abstract}

\section{1. Introduction}

A two-stage design for a bioequivalence assessment is a valuable tool when little is known about the variability of the measured Pharmacokinetic (PK) parameters and the true difference between the two formulations for which bioequivalence must be demonstrated. There is an extensive literature on the use of two-stage group-sequential or adaptive designs in such a setting, but rather few examples of the use of these methods in practice (Kaza, Sokolovskyi, and Rudzki 2020). Specifically, confirmatory procedures that apply an adaptive sample size re-estimation after stage 1 have been suggested in for instance Maurer, Jones, and Ying (2018) and in practice several flexible software platforms enabling implementation of adaptive trials in the context of bioequivalence assessment are already established. However, even though all the tools to perform an adaptive bioequivalence trial are readily available a number of properties ensuring proper performance are still unclear. To fill this gap we provide some theoretical properties in the context of:

1. Implementation of a binding futility bound, aiming to prevent fruitless trials from continuing after the first stage,
2. Producing final confidence intervals that combine information from both stages and are consistent with the decision making,
3. The ability to efficiently handle multiple endpoints in the context of assessing simultaneous average bioequivalence.

Specifically, this paper considers a confirmatory adaptive Two One-Sided Tests (TOST) procedure with a built in futility bound, adapted to handle simultaneous bioequivalence of multiple pharmacokinetics endpoints, and accompanied by an overall confidence region that is ensured to support the final bioequivalence decision.

Following the above strategy a number of desirable properties are not automatically ensured and need further scrutiny. First, it is not per default ensured that the upper limit of the suggested overall confidence interval is in fact higher than the lower limit. In this article we show that under some mild assumptions this does in fact hold true.

Second, the traditional approach for assessing the simultaneous bioequivalence by using the maximal \( p \)-value to test the union of null-hypotheses against the intersection of the alternative hypotheses does not use the correlation between \( p \)-values and may therefore provide conservative control of the Type I error. We provide an alternative procedure for simultaneously assessing bioequivalence and show that this procedure not only provides Type I error control but is also more powerful than the traditional Intersection Union approach (Berger and Hsu 1996).

In a traditional fixed design setting the two one-sided test (TOST) procedure for assessing bioequivalence considers testing the null hypothesis \( H_0 \) that the targeted population difference \( \theta \) due to treatment is larger numerically than some equivalencemargin \( \Delta \), that is,

\[ H_0 : |\theta| > \Delta. \]

Due to the make-up of this null hypothesis as the union of the two one-sided null-hypotheses, say \( H_0^- : (\theta < -\Delta) \) and \( H_0^+ : (\theta > \Delta) \), a decision to reject the null hypothesis is equivalent to rejecting both one-sided null-hypotheses. Accordingly
the null-hypothesis is typically evaluated by considering two one-sided tests (TOST) for evaluating the accompanying two one-sided null hypotheses. The two test statistics used for this evaluation are typically standardized Z-scores measuring the standardized difference from the estimated targeted treatment difference to \(-\Delta\) and \(\Delta\), respectively.

Next \(p\)-values are calculated based on assuming that Z-score statistics are standard normal under the one-sided null hypotheses. Finally, The null hypothesis of non-bioequivalence is rejected at significance level \(\alpha\) if both of the resulting \(p\)-values are less than \(\alpha\).

In this classical setting a two-sided confidence interval with confidence level \(1 - 2\alpha\) is traditionally constructed assuming that the standardized difference between the estimated and true treatment difference follows a standard normal distribution based on referring the \(1 - 2\alpha\) normal area of the standard normal distribution back to an interval around the estimated treatment difference in which we are \(1 - 2\alpha\) confident to find the true treatment difference.

An alternative way of deriving the same confidence interval in this setting is based on the uniform distribution of the so-called shifted \(p\)-values (Wassmer and Brannath 2016, chap. 8).

This \(p\)-value based approach becomes particularly attractive in the adaptive setting we consider, where we only have the uniform distribution of the shifted overall \(p\)-values from a combination test to rely on (Brannath, Posch, and Bauer 2002). Moreover, by construction, the \(p\)-value based confidence interval is aligned to the decision regarding the two one-sided hypotheses as the lower limit of the resulting interval is above \(-\Delta\) if and only if the corresponding one-sided null-hypothesis \(H_0^+\) is rejected. Similarly, the upper limit is below \(\Delta\) if and only if \(H_0^-\) is rejected.

In Section 2 we describe the two stage confirmatory TOST procedure to test the two one-sided hypotheses based on a general combination test procedure. The resulting overall shifted \(p\)-values from that procedure are then used to develop an overall \(1 - 2\alpha\) confidence interval for the targeted treatment difference \(\theta\) that aligns with the final decision about rejection/acceptance of the two one-sided null-hypotheses. The resulting procedure is then extended to handle two target parameters simultaneously in Section 3. Next, in Section 4, we outline a sample size reestimation procedure intended to ensure a given overall power. All developments are then combined in Section 5 where we present simulation results to support the planning of a bioequivalence study comparing two different administrations of a antibody treatment of moderate to severe atopic dermatitis. Finally, Section 6 contains a discussion of the developed methods.

2. An Adaptive TOST Procedure

To fix notation, let the stage wise test statistics for stage \(j = 1, 2\) be defined by

\[
Z_{j-} = e_j(\theta) + \frac{\theta + \Delta}{\sigma_{nj}}, \quad Z_{j+} = -e_j(\theta) + \frac{-\theta + \Delta}{\sigma_{nj}}
\]

where \(e_1(\theta) \perp e_2(\theta)\) and \(e_j(\theta) = \frac{\hat{\theta} - \theta}{\sigma_{nj}} \sim N(0, 1)\) and \(\sigma_{nj} = \sigma / \sqrt{n_j}\) with \(\sigma\) denoting an appropriate dispersion parameter

and \(n_j\) denoting the sample size at stage \(j\). Large values of the tests are considered critical and the resulting stagewise \(p\)-values are given as

\[
p_{j-} = 1 - \Phi(Z_{j-}), \quad p_{j+} = 1 - \Phi(Z_{j+}).
\]

For an efficacy bound \(\alpha_1\), futility bound \(\alpha_0\), and a corresponding combination function \(C\), the overall \(p\)-values are defined as in (Wassmer and Brannath 2016, chap. 8). Specifically, the overall \(p\)-values \(Q(p_{1-}, p_{2-}, \alpha_1, \alpha_0)\) and \(Q(p_{1+}, p_{2+}, \alpha_1, \alpha_0)\) are given by

\[
Q(p_{1-}, p_{2-}, \alpha_1, \alpha_0) = p_1 \cdot I(p_{1-} \leq \alpha_1) + I(p_{1-} > \alpha_1) + I(\alpha_0 > p_{1-} > \alpha_1)
\]

\[
\cdot \left[\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 I(C(x, y) \leq C(p_{1-}, p_{2-}))dydx\right]
\]

\[
= p_1 \cdot I(p_{1-} < \alpha_1) + I(p_{1-} > \alpha_0) + I(\alpha_0 > p_{1-} \geq \alpha_1)
\]

\[
\cdot \left[\alpha_1 + P(\alpha_0 > X \geq \alpha_1, C(X, Y) \leq C(p_{1-}, p_{2-}))\right]
\]

for \(X\) and \(Y\) independent and uniformly distributed.

The null hypothesis of non-bioequivalence is then rejected at significance level \(\alpha\) according to the union-intersection principle Berger and Hsu (1996), that is, if

\[
\max(Q(p_{1-}, p_{2-}, \alpha_1, \alpha_0), Q(p_{1+}, p_{2+}, \alpha_1, \alpha_0)) < \alpha
\]

Notice that this decision rule is different from the one described in Maurer, Jones, and Ying (2018). There, if one proceeds to stage 2, both hypotheses \(H_0^+\) and \(H_0^-\) are reevaluated even if one of them is formally rejected at stage 1. In their procedure it is only possible to stop after stage 1 if both hypotheses are rejected simultaneously. Their procedure is confirmatory as efficacy bounds at stages 1 and 2 are chosen to control the overall Type I error but the potential reevaluation of a rejected null-hypothesis may lead to increased Type II error.

Moreover, futility is not actively built into their procedure. However, the procedure will still control Type I error irrespective of whether or not futility after stage 1 is enforced (Wassmer and Brannath 2016, sec. 2.3). For the suggestion in this article there is the option to actively incorporate planned stopping for futility in the decision procedure (\(\alpha_0 < 1\)). In that case, reevaluating a null-hypothesis at stage 2 that has exceeded the binding futility bound at stage 1 and thus disregarding the built in futility stop, would clearly inflate the Type I error rate (Wassmer and Brannath 2016, sec. 6.2.1). The suggested procedure may, however, also be applied without this feature by using \(\alpha_0 = 1\) and will in that case also provide Type I error control irrespective of whether stopping for futility is enforced at stage 1 (Wassmer and Brannath 2016, sec. 2.3).

In terms of power our suggestion has the advantage that if we reject one of the one-sided hypotheses already at stage 1 but still proceed to stage 2, we don’t have to reject that same null-hypothesis again after stage 2. To this end note that our suggestion also requires choosing the stage 1 futility and efficacy bounds, the choice of which will impact power to declare bioequivalence. There is a substantial literature on how to choose these quantities in order to optimize operating characteristics of the design especially in group sequential trials (see for instance Jennison and Turnbull 2000 or Wassmer and Brannath
2016). We do not pursue this further here but mention it for completeness.

To provide a two-sided confidence interval with confidence level 1−2α we consider the following shifted stagewise p-values

\[ p_j^\delta = 1 - \Phi(\varepsilon_j(\delta)), \quad p_j^{\delta+} = 1 - \Phi(-\varepsilon_j(-\delta)). \]

Testing the hypotheses:

\[ H_0^\delta : \theta \leq \delta, \quad H_0^{\delta+} : -\theta \leq \delta. \]

Furthermore define

\[ \alpha_1^\delta = 1 - \Phi(\Phi^{-1}(1 - \alpha_1) - \frac{\delta + \Delta}{\sigma_{1n}}), \]
\[ \alpha_0^\delta = 1 - \Phi(\Phi^{-1}(1 - \alpha_0) - \frac{\Delta + \delta}{\sigma_{1n}}). \]

Finally define the overall shifted p-values as

\[ p^\delta = Q(p_{1-1}^\delta - p_{2-1}^\delta, \alpha_1^\delta), \quad p^{\delta+} = Q(p_{1+1}^\delta + p_{2+1}^\delta, \alpha_1^\delta, \alpha_0^\delta). \]

As before we identify the lower bound \( l \) as the smallest \( \delta \) such that

\[ p_{1-}^\delta \geq \alpha, \quad (1) \]

and the upper bound \( u \) is identified as \( -\delta \) with the smallest \( \delta \) such that

\[ p_{1+}^\delta \geq \alpha. \quad (2) \]

Note that the above definition of the two sided overall confidence interval is aligned to the actual declaration of equivalence in the following manner:

\[ p_{1-}^\Delta = p_{1-}^{\delta-\Delta}, \quad p_{1+}^\Delta = p_{1+}^{\delta+\Delta} \]

so that nonequivalence is not rejected at level \( \alpha/2 \) exactly if either \( -\Delta \) or \( \Delta \) or both belong to the constructed two-sided \( 1 - \alpha \) confidence interval.

In a trial with a large enough stage 1 sample we also have that \( l < u \). This is summarized in Proposition 1 the proof of which is presented in Appendix A.1.

**Proposition 1.** For \( \alpha_1 < \alpha \) and \( \alpha_1 < \alpha_0 < 1 \) such that \( 2Z_{1-\alpha} - Z_{1-\alpha_1} + Z_{1-\alpha_0} > 0 \), and for \( \frac{\Delta}{\sigma_{1n}} > Z_{1-\alpha_1} - Z_{1-\alpha} \) it holds that \( l < u \).

We note that the assumption \( 2Z_{1-\alpha} - Z_{1-\alpha_1} + Z_{1-\alpha_0} > 0 \) restricts the combinations of efficacy and futility bounds that can be used in accordance with Proposition 1 (see Figure 1 for an illustration) to ensure a meaningful confidence interval aligned to the decision on equivalence. Intuitively, the restriction is needed to ensure that stage 1 and 2 point estimates cannot differ too dramatically as this would compromise the validity of the confidence interval if one bound was based only on stage 1 data and the other on data from both stages. Practically, this means that high futility bounds are not supported by Proposition 1. In particular, the results of Proposition 1 are not applicable if you continue to stage 2 unless equivalence is declared at stage 1 (\( \alpha_0 = 1 \)). However, as illustrated in Figure 1 with \( \alpha = 0.05 \), any futility bound \( \alpha_0 \) below 0.95 is feasible and for futility bounds below 0.5 no real restriction is put on the efficacy bound as the requirement in that case is \( \alpha_1 > 0.0005 \).

To this end note that specifically the formal decision procedure prompts that you may reach a decision not in favor of bioequivalence or partially in favor of bioequivalence at stage 1 but still go on to stage 2. For instance you could decide in favor of not rejecting or rejecting \( H_0^\delta \) at stage 1, but still go on to stage 2 to decide whether to reject \( H_0^{\delta+} \). In this case the adaptive design, if followed as described, does not use stage 2 data for a reevaluation of \( H_0^\delta \).

One could rightfully argue that the strategy of continuing to stage 2 when you have already given up on declaring bioequivalence is fruitless. However, on the other hand, it may still be of value to learn if either noninferiority or nonsuperiority may be declared in the absence of the other, for instance if you wanted to adjust the dosing of your test formulation or test device. To this end also notice that if stage 1 is designed to be large enough to ensure that \( p_{1-} > \alpha_0 \) dictates \( p_{1+} < \alpha_1 \) or vice versa we would always stop at stage 1 if we give up on declaring bioequivalence at that stage. This would be the case if we could ensure that \( \frac{\Delta}{\sigma_{1n}} > Z_{1-\alpha_1} + Z_{1-\alpha_0} \).
Finally, note that according to Proposition A.1 in Brannath, Posch, and Bauer (2002) the shifted \( p \)-values are uniformly distributed for \( \delta = \theta \). Consequently \( \theta \) will be below \( i \) with probability \( \alpha \) and above \( u \) with probability \( \alpha \). In the above exposition this of course relies on us knowing the true value of the dispersion parameter \( \sigma \). This would never be the case in practice and we would replace \( Z_{l-} \) and \( Z_{l+} \) by the corresponding standard \( t \)-statistics:

\[
T_{j-} = \frac{\sqrt{n_j}(\hat{\theta}_j + \Delta)}{\hat{\sigma}_j}, \quad T_{j+} = \frac{\sqrt{n_j}(-\hat{\theta}_j + \Delta)}{\hat{\sigma}_j},
\]

where \( \hat{\sigma}_j \) is an estimator of \( \sigma \) based on \( j \)th stage data. This has the consequence that when applied as is the adaptive TOST only ensures Type I error control and correct coverage of confidence intervals asymptotically. It is possible to improve small sample behavior by making assumptions about the exact simultaneous distribution of \((\hat{\theta}_j, \hat{\sigma}_j)\), but we do not pursue this further here. The performance of the proposed adaptive TOST procedure based on standard \( t \)-statistics is evaluated by simulation in Section 5.

### 3. Multiple Endpoints

In a PK bioequivalence trial focus is predominantly on establishing equivalence with respect to both log of the maximal concentration \( \log(C_{\text{max}}) \) and log of the area under the concentration curve \( \log(AUC) \) as required by regulatory authorities U.S. Food and Drug Administration Draft Guidance (2021). The traditional approach is to power the study to one of them Maurer, Jones, and Ying (2018). The rationale for such an approach is that typically the correlation between \( \log(C_{\text{max}}) \) and \( \log(AUC) \) will be very high and also the target parameter \( \theta^{(1)} \) for \( \log(C_{\text{max}}) \) will be approximately equal to the target parameter \( \theta^{(2)} \) for \( \log(AUC) \). As a consequence any assessment with respect to bioequivalence based on \( \log(C_{\text{max}}) \) is expected to be mirrored in a similar assessment based on \( \log(AUC) \). Clearly such a strategy is vulnerable even if the rationale is due to some lack of Type I error control. An excellent and in detail discussion of these challenges is available in Pallmann and Jaki (2017).

We propose a more direct approach based on assessing the target parameters \( \min(\theta^{(1)}, \theta^{(2)}) \) and \( \max(\theta^{(1)}, \theta^{(2)}) \). Specifically we propose to test the following hypotheses:

\[
H_0^{k-} \cup H_0^{k+} : \min(\theta^{(1)}, \theta^{(2)}) < -\Delta, \\
H_0^{k+} \cup H_0^{k-} : \max(\theta^{(1)}, \theta^{(2)}) > \Delta
\]

where \( H_0^{k-} \) and \( H_0^{k+} \) denote \( k \)th endpoint hypotheses in scope, that is,

\[
H_0^{k-} : \theta^{(k)} < -\Delta, \\
H_0^{k+} : \theta^{(k)} > \Delta.
\]

In order to do so define the \( j \)th stage estimator of \( \min(\theta^{(1)}, \theta^{(2)}) \) and \( \max(\theta^{(1)}, \theta^{(2)}) \) as:

\[
\hat{\theta}_j^{\min} = A_j\hat{\theta}_j^{(1)} + (1 - A_j)\hat{\theta}_j^{(2)}, \\
\hat{\theta}_j^{\max} = A_j\hat{\theta}_j^{(2)} + (1 - A_j)\hat{\theta}_j^{(1)},
\]

where \( A_j \) equals one if \( \hat{\theta}_j^{(1)} \leq \hat{\theta}_j^{(2)} \) and zero otherwise. Moreover define:

\[
\sigma_j^{\min} = A_j\sigma_{j}^{(1)} + (1 - A_j)\sigma_{j}^{(2)}, \\
\sigma_j^{\max} = A_j\sigma_{j}^{(2)} + (1 - A_j)\sigma_{j}^{(1)}.
\]

Finally define the stagewise tests and \( p \)-values as

\[
Z_{j-}^{\min} = \frac{\hat{\theta}_j^{\min} + \Delta}{\sigma_j^{\min}}, \quad Z_{j+}^{\max} = \frac{\hat{\theta}_j^{\max} + \Delta}{\sigma_j^{\max}},
\]

\[
Z_{j-}^{\min} = A_jZ_{j-}^{(1)} + (1 - A_j)Z_{j-}^{(2)}, \\
Z_{j+}^{\max} = A_jZ_{j+}^{(2)} + (1 - A_j)Z_{j+}^{(1)}.
\]

\[
p_j^{\min} = 1 - \Phi(Z_{j-}^{\min}) = A_jp_{j-}^{(1)} + (1 - A_j)p_{j-}^{(2)}, \quad p_j^{\max} = 1 - \Phi(Z_{j+}^{\max}) = A_jp_{j+}^{(2)} + (1 - A_j)p_{j+}^{(1)}.
\]

where \( Z_{j-}^{(k)} Z_{j+}^{(k)}, p_{j+}^{(k)} \) denote previously defined stagewise test statistics and \( p \)-values for the \( k \)th endpoint.

With these definitions \( H_0^{1-} \cup H_0^{2-} \) is assessed as in Section 2 using the stagewise \( p \)-values \( p_j^{\min} \) and likewise \( H_0^{1+} \cup H_0^{2+} \) is assessed using the stagewise \( p \)-values \( p_j^{\max} \). The overall \( p \)-values of these assessments may be calculated as \( Q(p_0^{\min}, p_1^{\min}, \alpha_1, \alpha_0) \) and \( Q(p_0^{\max}, p_1^{\max}, \alpha_1, \alpha_0) \), respectively. Also mimicking the developments in Section 2 the shifted stagewise \( p \)-values may be defined as

\[
p_j^{\min} = 1 - \Phi(e_j^{\min}(-\delta)), \quad p_j^{\max} = 1 - \Phi(-e_j^{\max}(-\delta)),
\]

where

\[
e_j^{\min}(-\delta) = \frac{\hat{\theta}_j^{\min} - \delta}{\sigma_j^{\min}}, \\
e_j^{\max}(-\delta) = \frac{\hat{\theta}_j^{\max} - \delta}{\sigma_j^{\max}}.
\]

From the shifted \( p \)-values a \( 1 - \alpha \) lower bound \( p_{\min}^{\min} \) for \( \min(\theta^{(1)}, \theta^{(2)}) \) as well as a \( 1 - \alpha \) upper bound \( p_{\max}^{\max} \) for \( \max(\theta^{(1)}, \theta^{(2)}) \) may be calculated as in Section 2. Below we present a result showing that under similar assumptions as those imposed in Proposition 1 the above procedure leads to a meaningful confidence interval. We first define the following quantities used in the assumptions

\[
\omega = \sigma_{1n}^{\min}/\sigma_{1n}^{\max}, \\
\kappa = Z_{1-\alpha_1} - \frac{2\Delta}{\sigma_{1n}^{\min} + \sigma_{1n}^{\max}}.
\]

### Proposition 2

Assume that \( \alpha_1 < \alpha \) and \( \alpha_1 < \alpha_0 < 1 \) is such that \( \min\{1 + \omega, 1 + 1/\omega\}Z_{1-\alpha} - Z_{1-\alpha_1} + Z_{1-\alpha_0} > 0 \). Also assume that:

\[
\Phi\left(-\omega \kappa + \frac{\sqrt{\kappa^2 + 2\log(w)\frac{w-1}{w+1}}}{w-1}\right) > 2\alpha.
\]
Then it holds that $\mu_{\text{min}} < \mu_{\text{max}}$.

Furthermore a sufficient condition for (3) to hold is that

$$-\kappa - \frac{\sqrt{k^2 + 2\log(w)\frac{w-1}{w+1}}}{2} > -Z_{1-\alpha}.$$ 

Note that compared to Proposition 1 the assumption $\min\{1+\omega, 1+1/\omega\}Z_{1-\alpha} - Z_{1-\alpha_1} + Z_{1-\alpha_0} > 0$ is stricter than the assumption $Z_{21-\alpha} - Z_{1-\alpha_1} + Z_{1-\alpha_0} > 0$. It requires that the precision with which we estimate the two endpoint parameters are not to different in the sense that $\sigma_{1n}^2 / \sigma_{1n}^2 < Z_{1-\alpha_1} - Z_{1-\alpha_0} - Z_{1-\alpha}$.

When $\sigma_{1n}^2 = \sigma_{1n}^2$ corresponding to $\omega = 1$ the first assumption of Proposition 2 reduces to the assumption $2Z_{1-\alpha} - Z_{1-\alpha_1} + Z_{1-\alpha_0} > 0$ in Proposition 1. Moreover, both (3) and the sufficient condition reduce to the assumption, $\Delta > Z_{1-\alpha_1} - Z_{1-\alpha_0}$ in Proposition 1. We give a formal proof of Proposition 2 in Appendix A.2.

For the above procedure to remain a valid procedure in terms of controlling the Type I error and to ensure coverage of the proposed confidence limits we need to show that the stagewise p-values defined above are asymptotically p-values. This is summarized in Proposition 3. A formal proof of Proposition 3 is presented in Appendix A.3.

**Proposition 3.** If $H_0^+ \cup H_0^-$ is true then for any $0 \leq \alpha \leq 1$:

$$\lim_{n_1 \to \infty} P(p_{\text{min}} \leq \alpha) \leq \alpha.$$

If $H_0^+ \cup H_0^+$ is true then for any $0 \leq \alpha \leq 1$:

$$\lim_{n_1 \to \infty} P(p_{\text{max}} \leq \alpha) \leq \alpha.$$

For a traditional approach we note that the adaptive procedure developed in Section 2 may be applied marginally to assess $H_0^-$ and $H_0^{+k}$ for each k separately. Clearly this requires that we proceed to stage 2 of the adaptive trial unless all 4 stage 1 p-values are either above the futility bound $\alpha_0$ or below $\alpha_1$. A final judgment about bioequivalence, that is assessing $H_0^+ \cup H_0^+ \cup H_0^- \cup H_0^-$ can then be made using the intersection-union principle. Specifically, if both the traditional approach $H_0^+ \cup H_0^-$ is rejected if max$_k\{Q(p_{1-\alpha-k}^\text{min}, 1-\alpha_0)\} < \alpha$ and $H_0^+ \cup H_0^+$ is rejected if max$_k\{Q(p_{1-\alpha-k}^\text{max}, 1-\alpha_0)\} < \alpha$. In comparison our suggestion rejects $H_0^+ \cup H_0^-$ if $Q(p_{1-\alpha-k}^\text{min}, 1-\alpha_0) < \alpha$ and $H_0^+ \cup H_0^+$ if $Q(p_{1-\alpha-k}^\text{max}, 1-\alpha_0) < \alpha$. By construction our suggestion is more powerful if $A_1 = A_2$ since then $Q(p_{1-\alpha-k}^\text{min}, 1-\alpha_0) \leq \max_k\{Q(p_{1-\alpha-k}^\text{min}, 1-\alpha_0)\}$ and $Q(p_{1-\alpha-k}^\text{max}, 1-\alpha_0) \leq \max_k\{Q(p_{1-\alpha-k}^\text{max}, 1-\alpha_0)\}$.

**4. Sample Size Reestimation**

The adaptive sample size algorithm proposed in Maurer, Jones, and Ying (2018) is tailored to their proposed decision rule, which rejects $H_0$ at the end of stage 1 if both $H_0^-$ and $H_0^+$ are rejected. Otherwise, based on their decision rule, the trial proceeds to stage 2, where both hypotheses, $H_0^-$ and $H_0^+$ are tested. As previously stated, our decision rule differs, as it allows for decisions regarding $H_0^-$ and $H_0^+$ to be taken at different stages. In this section, we develop a sample size reassessment algorithm specific to this proposed decision strategy.

Similar to the SSR algorithm in Maurer, Jones, and Ying (2018), the stage 2 sample size, $N_2$ is selected in order to ensure an overall target power of $1 - \beta$ for declaring bio-equivalence, under the alternative hypothesis corresponding to the interim estimates of $\theta$ and $\sigma$. If we denote the estimated stage 1 power under this alternative hypothesis by

$$1 - \beta_1 = P(\hat{\theta}_{1|1}, \theta_{1|1}(\alpha_1, p_{1+} \leq \alpha_1),$$

then the minimum required conditional power, $\hat{\theta}_{1|1}(p_{1-}, p_{1+})$, in order to ensure an overall target power of $1 - \beta$ for declaring bio-equivalence is given by

$$\frac{1 - \beta_1}{1 - \beta_0} = \frac{\beta_1 - \beta}{\beta_1 - \beta_0},$$

where

$$1 - \beta_0 = P(\hat{\theta}_{1|1}, p_{1-} < \alpha_0, p_{1+} < \alpha_0)$$

is the probability the trial does not stop due to futility at stage 1, under the specified alternative. The conditional power, $\hat{\theta}_{1|1}(p_{1-}, p_{1+})$, will take different forms depending on the stage 1 outcome. If at the end of stage 1, $p_{1-} < \alpha_0$ and $p_{1+} < \alpha_0$, then the conditional power will be given by one of the following four scenarios,

$$\hat{\theta}_{1|1}(p_{1-}, p_{1+}) = \begin{cases} 1, & p_{1-} < \alpha_1 \text{ and } p_{1+} \leq \alpha_1, \\ p_{1-} < \alpha_1 \text{ and } p_{1+} < \alpha_1, \\ p_{1-} < \alpha_1 \text{ and } p_{1+} < \alpha_0, & p_{1-} < \alpha_1 \text{ and } p_{1+} < \alpha_0, \\ p_{1-} < \alpha_1 \text{ and } p_{1+} < \alpha_0, & p_{1-} < \alpha_1 \text{ and } p_{1+} < \alpha_0. \end{cases}$$

Here, $A(\alpha)$ denotes the conditional error function associated with a specific combination function $C$, (see Wassmer and Brannath (2016), sec. 6.3). If we denote the desired stage 2 conditional power by $\alpha_{p_0}$, then under scenario 2, where $\alpha_1 < p_{1-} < \alpha_0$ and $p_{1+} \leq \alpha_1$, the required stage 2 samples size per arm, assuming 1:1 randomization, is given by

$$N_2 = \frac{2\alpha_{1|1}(1 - A(p_{1+}))) - \Phi^{-1}(1 - \alpha_{1|1}))^2}{(\Delta + \hat{\theta}_{1|1})},$$

which follows the derivation given in Wassmer and Brannath (2016). Similarly, for scenario 3, the required stage 2 sample size per arm would be given by

$$N_2 = \frac{2\alpha_{1|1}(1 - A(p_{1+}))) - \Phi^{-1}(1 - \alpha_{1|1}))^2}{(\Delta - \hat{\theta}_{1|1})^2}.$$
Wassmer and Brannath (2016) provides an in-depth overview of the potential issues that may arise for sample size re-estimation algorithms on based conditional power arguments using the interim parameter estimates. Briefly, the above formulas for the stage 2 sample size, are increasing functions in \( p_1 \). As the stage 1 \( p \)-value tends to 1, the required stage 2 sample size may balloon. Additionally, values of \((\Delta + \hat{\theta}_1)\) and \((\Delta - \hat{\theta}_1)\) close to zero would also cause the required stage 2 sample size to balloon. This may lead to designs with poor operating characteristics, for example large efficiency losses compared to standard trials design or group-sequential designs with similar expected sample sizes. For relative stage two samples sizes, \( N_2/N_1 \) less than 4, Wassmer and Brannath (2016) has demonstrated that the relative efficiency loss of the adaptive design compared to a standard design is minimal. Such situations, where the relative stage 2 sample size exceeds 4 may be avoided by either specifying a maximum stage 2 sample size, or through judicious choice of the stage 1 futility bound, \( \alpha_0 \).

So far we have neglected to address the case in which the \( p \)-value associated with one of the hypotheses, exceeds the futility bound at the interim analysis while the remaining \( p \)-value lies within the interval, \((\alpha_1, \alpha_0)\). In such an instance, one may benefit from increasing the precision of the estimated treatment difference and, or demonstrating noninferiority, with respect to the hypothesis to be tested in stage 2.

We will motivate the sample size re-estimation algorithm for this case in a similar vein to the above algorithm, in that we wish to select a stage 2 sample size, \( N_2 \), in order to ensure an overall target power of 1 - \( \gamma \), under the alternative hypothesis corresponding to the interim estimates of \( \theta \) and \( \sigma \). Specifically the targeted power corresponds to the probability of rejecting exactly one hypothesis given that exactly one stage 1 \( p \)-value exceeds the futility bound, that is

\[
1 - \gamma = 1 - \gamma_1 = \frac{1 - \gamma_1 + \int_{\gamma_2} \mathcal{C} \mathcal{P}_{\bar{\theta}_1, \hat{\gamma}_1}(p_1, p_{1+}) d \mathcal{P}_{\bar{\theta}_1, \hat{\gamma}_1}(p_1, p_{1+})}{1 - \gamma_0},
\]

where, \( S \) denotes the set, \( \{p_1 \in (\alpha_1, \alpha_0), p_{1+} \geq \alpha_0 \} \cup \{p_1 \geq \alpha_0, p_{1+} \in (\alpha_1, \alpha_0)\} \). The minimum required conditional power, \( \mathcal{C} \mathcal{P}_{\bar{\theta}_1, \hat{\gamma}_1}(p_1, p_{1+}) \), required in order to ensure a target power of 1 - \( \gamma \) to reject the one remaining hypothesis is given by

\[
\frac{(1 - \gamma) \cdot (1 - \gamma_0) - 1 + \gamma_1 - \gamma_0}{1 - \gamma_0}.
\]

In this setting, the conditional power may take one of two forms,

\[
\mathcal{C} \mathcal{P}_{\bar{\theta}_1, \hat{\gamma}_1}(p_1, p_{1+}) = \begin{cases} \{ p_{2-} \leq A(p_{1-}), \quad \alpha_1 \leq p_{1-} < \alpha_0 \text{ and } p_{1+} \geq \alpha_0 \\ p_{2+} \leq A(p_{1+}), \quad p_{1-} \geq \alpha_0 \text{ and } \alpha_1 \leq p_{1+} < \alpha_0 \} \end{cases}
\]

Under the assumption of 1:1 randomization, the minimum stage two sample size, \( N_2 \), can be obtained using the formulas for the stage 2 sample size derived above. The only difference being the value of the conditional power that is plugged into the sample size formulas. For further details regarding the sample size procedures, please refer to the discussion in Maurer, Jones, and Ying (2018). We also refer to the excellent discussion in Maurer, Jones, and Ying (2018) on how one may choose values of \( \theta \) and \( \sigma \) based on stage 1 data to get a more informed sample size re estimation.

When two target parameters are assessed for bioequivalence simultaneously as in Section 3 we, however, need to extend the standard sample size procedures slightly. Below we describe the main points of the extension. Our approach is simulation based and based on the assumption that \((\hat{\theta}_j^{(1)}, \hat{\theta}_j^{(2)})\) follows a bivariate normal distribution (at least asymptotically) with mean \((\theta^{(1)}, \theta^{(2)})\) and variance \( \frac{1}{\gamma} \Sigma \). The covariance matrix \( \Sigma \) is in turn defined by the within subject variance/covariance of the \( \log(C_{\max}) \) and \( \log(AUC) \) measurements. For instance consider a two arm parallel or cross-over design with 1:1 randomization of treatments. In such a design let \( \log(C_{\max})_{ij} \) and \( \log(AUC)_{ij} \) denote the measurements for the \( i \)th subject with the \( j \)th treatment \( l = 1, 2 \). With this notation and assuming no missing measurements we have

\[
\Sigma = A \cdot \Gamma \cdot A^T,
\]

where \( \Gamma = \text{var} \{ \log(C_{\max})_{11}, \log(AUC)_{1i}, \log(C_{\max})_{22}, \log(AUC)_{i2} \} \) and

\[
A = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}.
\]

In a cross-over study the empirical variance matrix estimating \( \Gamma \) follows a Wishart distribution with \( n_j - 1 \) degrees of freedom. For a two arm parallel study the empirical within treatment arm variance matrices estimating the \( 2 \times 2 \) block diagonal matrices of \( \Gamma \) are independent and both follow Wishart distributions with \( n_j/2 - 1 \) degrees of freedom.

Consequently, We may simulate realizations of \((\hat{\theta}_j^{(1)}, \hat{\theta}_j^{(2)})\) based on their multivariate normal distribution. We may furthermore simulate realizations of \( \hat{\Sigma} = A \cdot \hat{\Gamma} \cdot A^T \) by simulating \( \hat{\Gamma} \).
as described above. By combining these quantities as described in Section 3 we may produce realizations of $Z_j^{\min}$, $Z_j^{\max}$ and in turn $p^{\min}_1$, $p^{\max}_1$.

The above recipe enables simulation based estimation of both stage 1 power probability:

$$P_{\theta, \sigma}(p^{\min}_1 < \alpha_1, p^{\max}_1 < \alpha_1)$$

and conditional power probabilities

$$c_{\theta, \sigma}^0(p^{\min}_1) = P_{\theta, \sigma}(p^{\min}_1 < A(p^{\min}_1) p^{\min}_1),$$
$$c_{\theta, \sigma}^1(p^{\max}_1) = P_{\theta, \sigma}(p^{\max}_1 < A(p^{\max}_1) p^{\max}_1),$$
$$c_{\theta, \sigma}^2(p^{\min}_1, p^{\max}_1) = P_{\theta, \sigma}(p^{\min}_1 < A(p^{\min}_1), p^{\max}_1 < A(p^{\max}_1), p^{\min}_1 p^{\max}_1).$$

The fundamental difference between the sample size re-estimation in Maurer, Jones, and Ying (2018) and the one proposed here lies in the different decision strategies. Specifically the probability of declaring bioequivalence with their decision strategy in a given alternative $\theta, \sigma$ is given by

$$P_{\theta, \sigma}(p^{\min}_1 < \alpha_1, p^{\max}_1 < \alpha_1)$$
$$+ \int_0^{\alpha_1} \int_0^{\alpha_1} c_{\theta, \sigma}^0(p^{\min}_1, p^{\max}_1) dP_{\theta, \sigma}(p^{\max}_1, p^{\min}_1)$$
$$+ \int_0^{\alpha_1} \int_0^{\alpha_1} c_{\theta, \sigma}^1(p^{\min}_1, p^{\max}_1) dP_{\theta, \sigma}(p^{\max}_1, p^{\min}_1)$$
$$+ \int_0^{\alpha_1} \int_0^{\alpha_1} c_{\theta, \sigma}^2(p^{\min}_1, p^{\max}_1) dP_{\theta, \sigma}(p^{\max}_1, p^{\min}_1).$$

Since $c_{\theta, \sigma}^2(p^{\min}_1, p^{\max}_1)$ is smaller than both $c_{\theta, \sigma}^0(p^{\min}_1)$ and $c_{\theta, \sigma}^1(p^{\max}_1)$ it is clear that the decision strategy in Maurer, Jones, and Ying (2018) will yield less power to declare bioequivalence than the one outlined in this article. We elaborate on this in the simulation studies presented in the next section.

The sample size re-assessment when one stops for futility w.r.t one of the hypotheses already at stage 1 may seem irrelevant in the context of Section 2 where only one parameter is assessed for bioequivalence. Here one would expect that the probability of rejecting the other hypothesis at stage 1 is very high. Specifically, we noted in Section 2 that this probability is essentially 1 when the stage 1 sample is reasonably sized. However, in the context of Section 3 with multiple endpoints, where the null hypotheses target multiple parameters and are no longer disjoint, such a scenario may very well be envisaged and cannot be discarded.

5. Simulation Performance

In this section we investigate the feasibility of our proposal for a two-stage adaptive design in the context of designing a parallel treatment arm trial for assessing the equivalence of two different modes of administration of a biologic for treatment of atopic dermatitis. The motivation for using an adaptive design such as the proposed is a lack of knowledge about the actual pharmacokinetic difference between the two modes of administration. Safe-guarding against this uncertainty up front in a traditional fixed design would lead to a very high sample-size. Also there is an up front acknowledged risk that the two modes of administration may not be bioequivalent. If this is the case a built in futility stop is judged to be desirable.

The specific performance is considered for for the robust combination test suggested by Maurer, Jones, and Ying (2018). We term this the Maurer max combination test in what follows. To judge performance we investigate the following features:

1. The ability of our procedure to achieve the sought power of declaring bioequivalence.
2. The coverage and general behavior of the suggested overall confidence interval.

These features are investigated both for one endpoint and two endpoints simultaneously.

Finally a direct comparison between the suggested sample size re-estimation procedure and the sample size re-estimation procedure outlined in Maurer, Jones, and Ying (2018) is made to gauge the consequence of the different decision strategies in terms of overall power.

Before we proceed to simulations, however, we give a short introduction to the Maurer max combination test with the purpose of bridging the exposition in Maurer, Jones, and Ying (2018) to Section 2.

5.1. The Maurer Max Combination Test

Following the notation of Maurer, Jones, and Ying (2018) we define $Z_{ij} = \Phi^{-1}(1 - p_j)$, where $p_j, j = 1, 2$ denote the stagewise $p$-values. The max Maurer combination test is then given by

$$Z_{\max} = \max[wZ_1 + \sqrt{1 - w^2}Z_2, w_\ast Z_1 + \sqrt{1 - w_\ast^2}Z_2],$$

where $w$ and $w_\ast$ are two prespecified weights between zero and one. Assuming that $Z_1$ and $Z_2$ are independent and standard normal the $p$-value of the max combination test is calculated as

$$P(Z_{\max} > z_{\max}) = 1 - P(Z_{\max} \leq z_{\max})$$
$$= 1 - P(wZ_1 + \sqrt{1 - w^2}Z_2 \leq z_{\max})$$
$$w_\ast Z_1 + \sqrt{1 - w_\ast^2}Z_2 \leq z_{\max}).$$

Accordingly, the combination function for the Maurer max combination test is given by

$$C_{w, w_\ast}(p_1, p_2) = 1 - F(\max(\varepsilon(w, p_1, p_2), \varepsilon(w_\ast, p_1, p_2)))$$

with

$$\varepsilon(w, p_1, p_2) = w\Phi^{-1}(1 - p_1) + \sqrt{1 - w^2}\Phi^{-1}(1 - p_2)$$

and where for two standard normal-variables $\varepsilon_w$ and $\varepsilon_{w_\ast}$ with correlation $w \cdot w_\ast + \sqrt{(1 - w^2)(1 - w_\ast^2)}$

$$F(x) = P(\varepsilon_w \leq x, \varepsilon_{w_\ast} \leq x).$$

Consequently, for another standard normal variable $\varepsilon_1$ with correlation $w$ and $w_\ast$ to $\varepsilon_w$ and $\varepsilon_{w_\ast}$, respectively, we may write the overall $p$-value as

$$Q(p_1, p_2, \alpha_1, \alpha_0) = \{p_1 I[p_1 \in (0, \alpha_1) \cup (\alpha_0, 1)]$$
$$+ I[p_1 \in (\alpha_1, \alpha_0)]\}[\alpha_1 + P(Z_{1-\alpha_0} \leq \varepsilon_1 \leq Z_{1-\alpha_1}, max(\varepsilon_w, \varepsilon_{w_\ast}) \geq \max(\varepsilon(w, p_1, p_2), \varepsilon(w_\ast, p_1, p_2)))].$$

Notice that for $w = w_\ast$ the max Maurer combination test corresponds to the ordinary inverse normal combination test.
5.2. Simulation Setup and Results

In our simulation setup we consider a parallel arm study with a stage 1 sample size of 40 in each arm. Initially we consider only one endpoint. Specifically, we simulate the log(Cmax) endpoint from a standard normal distribution with mean zero in the reference arm and mean \( \theta = \log(0.87), \log(0.95), \log(1) \) in the test arm. For the residual error a standard deviation of 0.294 corresponding to a coefficient of variation of 0.3 is applied in both arms. The bioequivalence threshold is set to \( \Delta = \log(1.25) \).

With an overall significance level \( \alpha = 0.05 \) and a prespecified weights \( w = \sqrt{0.5} \) and \( w_0 = 0.5, w_\alpha = \sqrt{0.5}, \) or \( w_\alpha = \sqrt{0.85} \) we consider the following combinations of efficacy and futility bounds: \((\alpha_1, \alpha_0) = (0.026, 1), (0.028, 0.5), (0.034, 0.2)\) for \( w_\alpha = 0.5, (\alpha_1, \alpha_0) = (0.030, 1), (0.031, 0.5), (0.034, 0.2)\) for \( w_\alpha = \sqrt{0.5}, \) and \((\alpha_1, \alpha_0) = (0.029, 1), (0.030, 0.5), (0.033, 0.5)\) for \( w_\alpha = \sqrt{0.85} \). Sample size reassessment was made based on stage 1 ML parameter estimates as described in Section 4 to ensure an overall power of 0.9 and with a maximal stage 2 sample-size of 300. Specifically this means that we are using \( 1 - \beta = 0.9 \). Conditional powers are estimated for each considered stage 2 sample size by simulation with 10,000 simulated two-sided tests.

For each parameter configuration we simulate 10,000 trials with sample size reassessment after stage 1 as described above if we proceed to stage 2. If we do not proceed to stage 2 the stage 2 sample size is set to zero. For each trial we calculate the 90% confidence limits in accordance with \( \alpha = 5\% \) as described in Section 2. The results are summarized in Table 1.

In optimistic scenarios lower futility seems to increase stage two sample size. For the pessimistic scenario \((\theta = \log(0.87))\) setting the futility according to \( \alpha_0 = 0.2 \) results in a notably lower stage 2 sample size but also a notably lower overall power than for \( \alpha_0 = 0.5 \). For \( \alpha_0 = 0.5 \) stage 2 sample sizes are marginally higher compared to \( \alpha_0 = 1 \) (no futility) for the very optimistic scenario \((\theta = \log(1))\), whereas for the pessimistic scenario they are lower. Overall powers are generally close to each other comparing \( \alpha_0 = 0.5 \) to \( \alpha_0 = 1 \). In terms of choice of \( w_\alpha \) the higher values \( \sqrt{0.5}, \sqrt{0.85} \) seem comparable and marginally better in terms of stage 2 sample size than \( w_\alpha = 0.5 \). As expected the stage 1 power increases marginally when we decrease the futility.

Turning to the confidence intervals acceptable coverage probabilities close to 5% are observed. When \( \alpha_0 = 1 \) Proposition 1 no longer guarantees that the lower bound is below the upper bound and in these scenarios we also observe instances with a lower bound above the upper bound. In accordance with Proposition 1 this is not observed when \( \alpha_0 \leq 0.5 \).

Our second simulation study is a direct extension of the scenarios considered in the first simulation. Here we also include the endpoint \( \log(\text{AUC}) \). We simulate \( \log(\text{AUC}) \) from the same marginal distribution as \( \log(C_{\text{max}}) \) and with a correlation between these to measurements of 0.8. Otherwise the setup in terms of effect size, \( w, w_\alpha, \alpha_0, \alpha_1, \) sample size reassessment,

### Table 1. Summary of results.

| \( \theta \) | \( \alpha_1 \) | \( \alpha_0 \) | \( u\text{C}l < \theta \) | \( \theta < u\text{C}l \) | \( u\text{C}l < |\text{C}l| \) | Avg. \( n_2 \) | Avg. power |
|-------------|---------------|---------------|-----------------|-----------------|-----------------|----------|----------|
| \( \log(1) \) | 0.026 | 1 | 0.048 | 0.050 | 0 | 23.950 | 1.000 | 0.863 |
|              | 0.028 | 0.5 | 0.050 | 0.051 | 0 | 28.673 | 0.999 | 0.863 |
|              | 0.034 | 0.2 | 0.053 | 0.051 | 0 | 32.400 | 0.989 | 0.881 |
| \( \log(0.95) \) | 0.026 | 1 | 0.055 | 0.051 | 0 | 53.287 | 0.996 | 0.733 |
|              | 0.028 | 0.5 | 0.052 | 0.052 | 0 | 56.299 | 0.994 | 0.746 |
|              | 0.034 | 0.2 | 0.049 | 0.050 | 0 | 56.970 | 0.964 | 0.774 |
| \( \log(0.87) \) | 0.026 | 1 | 0.055 | 0.051 | 0.0008 | 189.607 | 0.788 | 0.255 |
|              | 0.028 | 0.5 | 0.052 | 0.053 | 0 | 167.092 | 0.754 | 0.266 |
|              | 0.034 | 0.2 | 0.052 | 0.052 | 0 | 113.910 | 0.618 | 0.283 |

For each parameter configuration we simulate 10,000 trials with sample size reassessment after stage 1 as described above if we proceed to stage 2. If we do not proceed to stage 2 the stage 2 sample size is set to zero. For each trial we calculate the 90% confidence limits in accordance with \( \alpha = 5\% \) as described in Section 2. The results are summarized in Table 1.
and number of simulations is identical to the first simulation study. Results are summarized in Table 2.

The results of our second $C_{\text{max}}$ and AUC endpoint simulation study largely mirror those obtained in the $C_{\text{max}}$ only endpoint simulation study with two major exceptions.

First, the coverage of our 90% confidence interval is well above 90% in all considered scenarios. This is in line with Proposition 3 which ensures that the used stagewise shifted $p$-values in the true parameter values are only asymptotically p-clud instead of being asymptotically uniformly distributed. This guarantees that asymptotically the actual coverage is at least as large as the nominal coverage, but not that the actual coverage approaches the nominal coverage asymptotically.

Second, We note that compared to the $C_{\text{max}}$ endpoint simulation study we need somewhat larger stage 2 sample sizes to achieve similar power to detect bioequivalence. Intuitively, this may be a bit surprising as the additional endpoint we have added is highly correlated to our initial endpoint with a correlation of 0.8, but it is in line with the observations made in Pallmann and Jaki (2017).

We investigate the impact of between endpoint correlation in more detail in a third simulation study. In this simulation study we restrict ourselves to only studying performance when the true effect for both endpoints is $\theta(1) = \theta(2) = \log(1)$, when $\alpha_0 = 0.5$, and when $w = w_\alpha = 0.5$. The standard deviation for $\log(C_{\text{max}})$ and $\log(AUC)$ is set to 0.294. Between endpoint correlations of 0.8, 0.9, 0.95, 0.99, and 1 are considered. Results of the simulations are given in Table 3.

Table 3 shows that stage 2 sample size decreases toward that of a single endpoint as correlation increases to 1. When the correlation is 1 results are as expected nearly identical to the single endpoint results in Table 1 for the scenario $\theta = \log(1)$, $\alpha_0 = 0.5$, $w_\alpha = \sqrt{0.5}$. We also note that almost perfect correlation between the two endpoints is needed to achieve a performance similar to that of a single endpoint.

In a final simulation we investigate the magnitude of the expected gain in stage two sample size with the adaptive TOST procedure proposed in this paper compared to the adaptive procedure proposed in Maurer, Jones, and Ying (2018).
consider a similar setup as in the first simulation study restricting ourselves to no stopping for futility ($\alpha_0 = 1$) for which the adaptive procedure in Maurer, Jones, and Ying (2018) was originally developed. In addition to this we also consider a scenario with a lower stage 1 sample size ($n_1 = 40$) and lower overall ensured power of 0.8. Finally, to eliminate Monte Carlo error, both procedures are based on the same simulated stage 1 data in each run of each scenario. Results are summarized in Table 4.

From Table 4 we note almost identical performance of the two procedures with a stage 1 sample size of 80. This is owing to the fact that the two hypotheses $H_0^1$ and $H_1^1$, if rejected at stage 1, are also rejected at stage 2 in virtually all simulations. In that case the two procedures coincide. When lowering the stage 1 sample size decisions to reject at stage 1 are contradicted at stage 2 more often and a minor gain in stage two sample size with the proposed adaptive TOST is observed. This is in line with the theoretical comparison provided in Section 4 but clearly of little practical relevance in the scenarios considered in this simulation study.

6. Discussion

In this article, we have presented a comprehensive two-stage adaptive TOST procedure for assessing bioequivalence. The approach, which incorporates formal stopping criteria, both in terms of futility and efficacy, at the interim analysis for both one-sided tests, has several advantages in comparison to previously published adaptive TOST procedures.

First and foremost, through the specification of the overall shifted $p$-values, confidence intervals, consistent with the decision of whether to reject the null hypothesis of nonequivalence, can be generated. It is not readily apparent how to go about generating such consistent confidence intervals when both hypotheses are retested when the trial continues to the second stage, for example methods B, C, and D of Potvin et al. (2008). Methods B, C, and D of Potvin et al. (2008) relied on informal stopping criteria at the interim analysis, for example, prespecified thresholds for the observed power to demonstrate average bioequivalence. However, the observed power is directly related to the $p$-values associated with the stage 1 TOST procedure. Depending on the intent of the first stage design, that is, pilot trial versus the sample size corresponding to a single stage study, the relevant threshold for the observed power may vary greatly and would necessitate the use of simulation to calibrate the threshold in order to ensure the proposed design has acceptable operating criteria. Our approach relies on the specification of formal stopping boundaries in terms of the stage 1 test statistics and hence the standard group sequential machinery can be applied when designing the trial.

The proposed framework also includes a formal algorithm for continuing to the second stage in the situation that one of the test statistics for the TOST procedure fails to exceed the futility bound at the interim. The proposed SSR algorithm selects the second stage sample size in order to ensure an overall target power conditional on exactly one hypothesis failing to reach the futility bound at the interim analysis. In practice, the method is implemented by deriving a bound for the conditional power that ensures the overall target power is met on average.

An often-overlooked aspect in designing a clinical trial to demonstrate the average bioequivalence of two drug formulations, is the regulatory requirement to demonstrate average bioequivalence with respect to both Cmax and AUC. It is common to power such a trial based on the requirements related to demonstrating average bioequivalence with regards to Cmax. In practice, this assumption is driven by the observed differences between the magnitudes of the variances for Cmax and AUC.
This assumption is, however, not inherently true. As a part of this paper, we have provided an efficient, simulation-based algorithm, to assess the operating characteristics of such designs. The algorithm derives its efficiency from simulating realizations from the joint distribution underlying the test statistics.

Our proposed implementation takes advantage of the pre-specified binding futility bound, by relaxing the threshold for rejecting the null hypothesis of nonequivalence. Additionally, crucial features of the proposed confidence intervals hinges on the use of a binding futility bound. However, based on the level of uncertainty at the design stage and/or the intent of the first stage of the design, for example, pilot study, it may be advantageous to either forgo the inclusion of a futility bound or specify the use of a nonbinding futility bound. For regulatory considerations regarding the use of binding/nonbinding futility bounds, please refer to the current FDA guidance on adaptive trial designs, U.S. Food and Drug Administration (2019). The flexibility of being able to continue to the second stage, while controlling the Type I error rate in the strong sense, despite failing to exceed the futility bound at the interim may outweigh the benefits of incorporating a binding bound, including the ability to derive sensible CIs.

We have chosen to use the combination function approach to control the Type I error rate in the strong sense. The examples and simulations presented in the article use the same combination functions for both one-sided hypothesis tests and Pocock type bounds for the efficacy stopping criteria at stage 1 and 2. These assumptions were driven by the design characteristics of the motivating example and ease of implementation. The theory presented is easily extended to accommodate combination functions and efficacy/futility bounds specific for each of the two hypotheses tested. For any given application, it may be beneficial to explore the impact of various combination functions, as well as futility and efficacy bounds for the stagewise test statistics.

To ensure the efficiency of our proposed sample size re-estimation algorithm based on the observed conditional power we introduced a limit on the maximum stage 2 sample size. A similar outcome could be achieved through a judicious choice of the stage 1 futility bound. On the other hand, sample size re-estimation algorithms using criteria other than the observed conditional power, for example, conditional power based on a pre-specified minimal, clinically relevant effect size, or predictive power, either in a Bayesian or frequentist setting, have been proposed in the literature. As there is no universally preferred method, the criteria underpinning the sample size re-estimation algorithm should be thoroughly assessed either analytically or via simulation. In terms of the motivating example presented in this article, given the relatively large stage 1 sample size, we would expect the SSR algorithm based on the observed conditional power to perform well, in terms of the expected sample size. This is due to the precision associated with the parameter estimates at the interim analysis. Recently, Hsiao, Liu, and Mehta (2019) published an optimal promising zone sample size re-estimation algorithm when testing for superiority. The approach is based on assessing the observed conditional power to reject the null hypothesis at the end of the stage 1, assuming the effect size is equal to the minimal clinically relevant effect size. It may be interesting to explore whether such an approach can be applied to the equivalence setting, by basing the SSR algorithm on the maximum effect size considered to be clinically equivalent.

Appendix

A.1. Proof of Proposition 1

We divide the proof into four scenarios

Scenario 1: \( p_i^{−Δ} \in (0, α_i) \cup (α_i, 1) \) and \( p_i^{+Δ} \in (0, α_i) \cup (α_i, 1) \)

Scenario 2: \( p_i^{−Δ} \in (α_i, α_0) \) and \( p_i^{+Δ} \in (α_i, α_0) \)

Scenario 3: \( p_i^{−Δ} \in (0, α_i) \cup (α_i, 1) \) and \( p_i^{+Δ} \in (α_i, α_0) \)

Scenario 4: \( p_i^{−Δ} \in (α_i, α_0) \) and \( p_i^{+Δ} \in (0, α_i) \cup (α_i, 1) \)

In Scenario 1 we have \( p_i^{−δ} = p_i^{1−} \) and \( p_i^{+δ} = p_i^{1+} \). Consequently solving (1) yields \( l = \hat{θ}_1 − σ_{1}(Z_{1−α}) \) and solving (2) yields \( u = \hat{θ}_1 + σ_{1}(Z_{1−α}) \). We conclude that \( l < u \) in Scenario 1.

For scenario 2 we note that

\[ p_i^{−δ} \geq α_i^{−} \] and \( p_i^{+δ} \geq α_i^{−} \)

and that for \( α < 0.5 \)

\[ α_i^{−} + α_i^{−} \geq I(Z_{1−α1} < \frac{Δ}{σ_{1α}}) \]

\[ + I(Z_{1−α1} ≥ \frac{Δ}{σ_{1α}} + 2 \cdot \Phi(\frac{Δ}{σ_{1α}} − Z_{1−α1}) > 2 \cdot α \]

The first inequality above follows from minimizing \( f(x, k) = \phi(x − k) \) as a function of \( x \) in the two scenarios \( k > 0 \) and \( k \leq 0 \). The second inequality follows from the assumption that \( \frac{Δ}{σ_{1α}} > Z_{1−α1} \).

This effectively means that since \( p_i^{1−} = α \) we have

\[ p_i^{1−} > α \]

From this we conclude that \( −u < −l \) in Scenario 2.

For Scenario 3 we first note that \( p_i^{1−} = p_i^{1−} \) so that \( l = \hat{θ}_1 − σ_{1α}(Z_{1−α}) \). We also note that the requirement \( p_i^{1−} < α_0 \) entails that \( Δ − \hat{θ}_1 > σ_{1α}(Z_{1−α}) \). Furthermore, solving

\[ α_i^{−} = α \]

yields

\[ u = σ_{1α}(Z_{1−α1} − Z_{1−α}) − Δ \]

Now since \( p_i^{+} > α_i^{−} \) we deduce that \( −u < −l \). Consequently

\[ u − l > Δ − \hat{θ}_1 + σ_{1α}(2 \cdot Z_{1−α} − Z_{1−α}) > 0 \]

where the last inequality follows from our assumptions.

Scenario 4 proceeds similarly to Scenario 3. We note that \( u = \hat{θ}_1 + σ_{1α}(Z_{1−α}) \) and that \( \hat{θ}_1 + Δ > σ_{1α}(Z_{1−α}) \) and finally that \( l < u \), where \( u \) is defined in Scenario 3. This entails that

\[ u − l > \hat{θ}_1 + Δ + σ_{1α}(2 \cdot Z_{1−α} − Z_{1−α}) > 0 \]

A.2. Proof of Proposition 2

Similar to the proof of Proposition 1 We consider four scenarios

Scenario 1: \( p_i^{−Δ\min} \in (0, α_i) \cup (α_i, 1) \) and \( p_i^{+Δ\min} \in (0, α_i) \cup (α_i, 1) \)

Scenario 2: \( p_i^{−Δ\min} \in (α_i, α_0) \) and \( p_i^{+Δ\min} \in (α_i, α_0) \)

Scenario 3: \( p_i^{−Δ\min} \in (0, α_i) \cup (α_i, 1) \) and \( p_i^{+Δ\min} \in (α_i, α_0) \)

Scenario 4: \( p_i^{−Δ\min} \in (α_i, α_0) \) and \( p_i^{+Δ\min} \in (0, α_i) \cup (α_i, 1) \)

In the first scenario we have \( p_i^{−\min} = p_i^{1−} \) and \( p_i^{+\max} = p_i^{1+} \). As a consequence we get \( \min = \frac{δ_{1−\min}}{\frac{\sigma_{1α}(Z_{1−α})}{\frac{\sigma_{1α}}{\frac{2α}{2α}}}} \) and \( \max = \frac{δ_{1+\max}}{\frac{\sigma_{1α}(Z_{1−α})}{\frac{\sigma_{1α}}{\frac{2α}{2α}}}} \).

Since \( \min ≤ 1 \) it follows that \( \min < \max \) in Scenario 1.

For Scenario 2 we note that

\[ p_i^{−\min} ≥ α_i^{−\min} \] and \( p_i^{+\max} ≥ α_i^{−\max} \)

and that for \( α < 0.5 \)
Then for any \( \sigma \) which follows (3), following \( k \) and the second inequality follows directly from assumption (3) in Proposition 2. As in Proposition 1 we conclude that \( -\mu_{\text{max}} < -\mu_{\text{min}} \) in Scenario 2.

For Scenario 3 we first note that \( P_{\mu_{\text{min}}} = P_{\mu_{\text{max}}} \) so that \( \mu_{\text{min}} = \mu_{\text{max}} = \sigma_{1n}^{\text{min}} \). We also note that the requirement \( P_{\mu_{\text{max}}} < \alpha_0 \) entails that \( \sigma_{1n}^{\text{min}} Z_{1-a0} < \Delta - \hat{\theta}_{1\text{max}} \leq \Delta - \hat{\theta}_{1\text{min}} \). Furthermore, solving

\[
\alpha_{1\text{max}} = \alpha
\]

yields

\[
\hat{\mu} = \sigma_{1n}^{\text{min}} (Z_{1-a1} - Z_{1-a0} - \Delta).
\]

Now since \( P_{\mu_{\text{max}}} > \alpha_{1\text{max}} \) we deduce that \( -\mu_{\text{max}} < \hat{\mu} \). Consequently

\[
u_{\text{max}} - \mu_{\text{min}} = \Delta - \hat{\mu}_{\text{min}} + (\sigma_{1n}^{\text{min}} + \sigma_{1n}^{\text{max}}) Z_{1-a0} - \sigma_{1n}^{\text{max}} Z_{1-a0} > 0
\]

where the last inequality follows from the assumption (1 + w) \( Z_{1-a0} - Z_{1-a1} - Z_{1-a0} > 0 \).

Scenario 4 proceeds similarly to Scenario 3. We note that \( u_{\text{max}} = \hat{\theta}_{1\text{max}} + \sigma_{1n}^{\text{max}} Z_{1-a0} \) and that \( \sigma_{1n}^{\text{min}} Z_{1-a0} < \hat{\mu}_{\text{min}} + \Delta \leq \hat{\mu}_{1\text{max}} + \Delta \). Furthermore, solving

\[
\alpha_{1\text{min}} = \alpha
\]

yields

\[
\hat{\mu} = \sigma_{1n}^{\text{min}} (Z_{1-a1} - Z_{1-a0} - \Delta).
\]

Now since \( P_{\mu_{\text{min}}} = \alpha_{1\text{min}} \) we deduce that \( \mu_{\text{min}} < \hat{\mu} \). Consequently

\[
u_{\text{max}} - \mu_{\text{min}} = \Delta + \hat{\theta}_{1\text{max}} + (\sigma_{1n}^{\text{min}} + \sigma_{1n}^{\text{max}}) Z_{1-a0} - \sigma_{1n}^{\text{min}} Z_{1-a0} > 0
\]

where the last inequality follows from the assumption (1 + w) \( Z_{1-a0} - Z_{1-a1} - Z_{1-a0} > 0 \).

Finally, to show that sufficient condition entails (3) first note that

\[
\rho = \frac{1}{2} \left( -\omega \alpha + \sqrt{k^2 + 2 \log(w) \frac{w-1}{w+1}} \right) + \frac{k - \omega \sqrt{k^2 + 2 \log(w) \frac{w-1}{w+1}}}{w-1} < 0.
\]

Then for any \( k > 0 \) we have \( \Phi(\rho - k) + \Phi(\rho + k) \geq 2 \Phi(\rho) > 2 \alpha \) from which (3) follows.

### A.3. Proof of Proposition 3

First we state following lemma that is crucial in proving the statement. The proof of the lemma follows from straightforward algebra.

**Lemma 1.** Let \( U, V \) be real numbers, put \( P = I(U \leq V) \), and let \( \sigma_1, \sigma_2 \) denote positive real numbers then the following holds:

\[
\frac{U \cdot P + V \cdot (1 - P) \cdot \sigma_2}{P \cdot \sigma_1 + (1 - P) \cdot \sigma_2} \leq \frac{U}{\sigma_1} I(\sigma_1 \leq \sigma_2) + \frac{V}{\sigma_2} I(\sigma_1 > \sigma_2),
\]

\[
\frac{U \cdot (1 - P) + V \cdot P}{(1 - P) \cdot \sigma_1 + P \cdot \sigma_2} \geq \frac{U}{\sigma_1} I(\sigma_1 > \sigma_2) + \frac{V}{\sigma_2} I(\sigma_1 \leq \sigma_2).
\]

For proving Proposition 3 first consider \( \theta(1) < \theta(2) \). In this scenario we have

\[
P(\hat{\theta}_j^{(1)} \geq \hat{\theta}_j^{(2)} \rightarrow 0 \text{ as } n_j \rightarrow \infty.
\]

This entails that

\[
Z_j = \frac{\hat{\theta}_j^{(1)} + \Delta}{\sigma_j^{(1)}} + \sigma_j^{(1)}.
\]

Accordingly when \( H_0^1 \cup H_0^2 \cap \{(\theta(1), \theta(2)) : \theta(1) < \theta(2)\} = H_0^1 \cap \{(\theta(1), \theta(2)) : \theta(1) < \theta(2)\} \) is true:

\[
\lim_{n_j \rightarrow \infty} \text{P}(\text{P}_{\text{min}}^{(1)} \leq \alpha) = \lim_{n_j \rightarrow \infty} \text{P}(\text{P}_{\text{min}}^{(2)} \leq \alpha) \leq \alpha.
\]

For the scenario \( \theta(1) > \theta(2) \) we similarly get:

\[
Z_j = \frac{\hat{\theta}_j^{(2)} + \Delta}{\sigma_j^{(2)}} + \sigma_j^{(2)}.
\]

Accordingly when \( H_0^1 \cup H_0^2 \cap \{(\theta(1), \theta(2)) : \theta(1) > \theta(2)\} = H_0^1 \cap \{(\theta(1), \theta(2)) : \theta(1) > \theta(2)\} \) is true:

\[
\lim_{n_j \rightarrow \infty} \text{P}(\text{P}_{\text{min}}^{(1)} \leq \alpha) = \lim_{n_j \rightarrow \infty} \text{P}(\text{P}_{\text{min}}^{(2)} \leq \alpha) \leq \alpha.
\]

Finally for the scenario \( \theta(1) = \theta(2) \) we use Lemma 1 to see that

\[
Z_j = \frac{\hat{\theta}_j^{(1)} + \Delta}{\sigma_j^{(1)}} + \sigma_j^{(1)} = \frac{\hat{\theta}_j^{(2)} + \Delta}{\sigma_j^{(2)}} + \sigma_j^{(2)}.
\]

Now let \( \sigma^{(k)} \) denote the limit in probability of \( \frac{1}{n_j} \mu^{(k)} \) for \( k = 1, 2 \) and \( j = 1, 2 \) and note that when \( \sigma(1) < \sigma(2) \) we may use the above equality to conclude that:

\[
\frac{\hat{\theta}_j^{(1)} + \Delta}{\sigma_j^{(1)}} + \sigma_j^{(1)} \leq \frac{\hat{\theta}_j^{(2)} + \Delta}{\sigma_j^{(2)}} + \sigma_j^{(2)} \text{I}(\sigma_j^{(1)} > \sigma_j^{(2)})
\]

\[
= \frac{\hat{\theta}_j^{(1)} + \Delta}{\sigma_j^{(1)}} + \sigma_j^{(1)} + \text{I}(\sigma_j^{(1)} > \sigma_j^{(2)})
\]

from which we conclude that when \( H_0^1 \cup H_0^2 \cap \{(\theta(1), \theta(2)) : \theta(1) = \theta(2)\} \) is true:

\[
\lim_{n_j \rightarrow \infty} \text{P}(\text{P}_{\text{min}}^{(1)} \leq \alpha) = \lim_{n_j \rightarrow \infty} \text{P}(\text{P}_{\text{min}}^{(2)} \leq \alpha) \leq \alpha.
\]

Similarly, when \( \sigma(1) > \sigma(2) \) we get that

\[
\frac{\hat{\theta}_j^{(1)} + \Delta}{\sigma_j^{(1)}} + \sigma_j^{(1)} \leq \frac{\hat{\theta}_j^{(2)} + \Delta}{\sigma_j^{(2)}} + \sigma_j^{(2)} \text{I}(\sigma_j^{(1)} > \sigma_j^{(2)})
\]

\[
= \frac{\hat{\theta}_j^{(1)} + \Delta}{\sigma_j^{(1)}} + \sigma_j^{(1)} + \text{I}(\sigma_j^{(1)} > \sigma_j^{(2)})
\]

from which we also conclude that when \( H_0^1 \cup H_0^2 \cap \{(\theta(1), \theta(2)) : \theta(1) = \theta(2)\} \) is true:

\[
\lim_{n_j \rightarrow \infty} \text{P}(\text{P}_{\text{min}}^{(1)} \leq \alpha) \leq \lim_{n_j \rightarrow \infty} \text{P}(\text{P}_{\text{min}}^{(2)} \leq \alpha) \leq \alpha.
\]

Finally for the case \( \sigma(1) = \sigma(2) \) note that:

\[
Z_j = \frac{\hat{\theta}_j^{(1)} + \Delta}{\min(\sigma_j^{(1)} , \sigma_j^{(2)})} + \sigma_j^{(1)} + \sigma_j^{(2)}.
\]

Thus, when \( H_0^1 \cup H_0^2 \cap \{(\theta(1), \theta(2)) : \theta(1) = \theta(2)\} \) is true, that is, when \( \theta(1) \leq \Delta \) it follows that:

\[
Z_j \leq \frac{\hat{\theta}_j^{(1)} - \theta(1)}{\sigma_j^{(1)}} + \sigma_j^{(1)} + \sigma_j^{(2)}.
\]
from which we conclude that
\[
\lim_{n_j \to \infty} P(p_j^\text{min} \leq \alpha) \leq \lim_{n_j \to \infty} P\left(\frac{\hat{\theta}_j^{(1)} - \theta_j^{(1)}}{\sigma_j^{(1)}} > Z_{1-\alpha}\right) = \alpha.
\]
The proof of the second statement in Proposition 3 follows along the same lines as above.

Acknowledgments

We would like to thank the editor and the two referees for their helpful comments.

ORCID

Christian Bressen Pipper http://orcid.org/0000-0003-0261-616X

References

Berger, R. L., and Hsu, J. C. (1996), "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets," Statistical Science, 11, 283–319. [192,193,196]

Brannath, W., Posch, M., and Bauer, P. (2002), "Recursive Combination Tests," Journal of the American Statistical Association, 97, 236–244. [195]

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2021), mvtnorm: Multivariate Normal and t Distributions. R package version 1.1-2. Available at https://CRAN.R-project.org/package=mvtnorm. [196]

Hsiao, S. T., Liu, L., and Mehta, C. R. (2019), "Optimal Promising Zone Designs," Biometrical Journal, 5, 1175–1186. [202]

Jennison, C., and Turnbull, B. (2000), Group Sequential Methods with Applications to Clinical Trials, Boca Raton, FL: Chapman & Hall/CRC Press. [193]

Kaza, M., Sokolovskyi, A., and Rudzki, P. J. (2020), "10th Anniversary of a Two-Stage Design In Bioequivalence. Why Has It Still Not Been Implemented?" Pharmaceutical Research, 37, 140. [192]

Labes, D., Lang, B., and Schütz, H. (2021), Power2Stage: Power and Sample-Size Distribution of 2-Stage Bioequivalence Studies. R package version 0.5-4. Available at https://CRAN.R-project.org/package=Power2Stage [196]

Maurer, W., Jones, B., and Ying, C. (2018), "Controlling the Type I Error Rate in Two-Stage Sequential Adaptive Designs When Testing for Average Bioequivalence," Statistics in Medicine, 37, 1587–1607. [192,193,195,196,197,198,200,201]

Pallmann, P., and Jaki, T. (2017), "Simultaneous Confidence Regions for Multivariate Bioequivalence," Statistics in Medicine, 36, 4585–4603. [195,200]

Patterson, S., and Jones, B. (2017), Bioequivalence and Statistics in Clinical Pharmacology (2nd ed.), Boca Raton, FL: CRC Press. [196]

Potvin, D., DiLiberti, C. E., Hauck, W. W., Parr, A. F., Schairmann, D. J., and Smith, R. A. (2008), "Sequential Design Approaches for Bioequivalence Studies with Crossover Designs," Pharmaceutical Statistics, 7, 245–262. [201]

U.S. Food and Drug Administration (2019), "Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry." [202]

U.S. Food and Drug Administration Draft Guidance (2021), "Bioequivalence Studies With Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA Guidance for Industry." [195]

Wassmer, G., and Brannath, W. (2016), Group Sequential and Confirmatory Adaptive Designs in Clinical Trials, Cham: Springer. [193,194,196,197]