Single-stage, three-arm, adaptive test strategies for non-inferiority trials with an unstable reference

Werner Brannath  |  Martin Scharpenberg  |  Sylvia Schmidt

For indications where only unstable reference treatments are available and use of placebo is ethically justified, three-arm “gold standard” designs with an experimental, reference and placebo arm are recommended for non-inferiority trials. In such designs, the demonstration of efficacy of the reference or experimental treatment is a requirement. They have the disadvantage that only little can be concluded from the trial if the reference fails to be efficacious. To overcome this, we investigate novel single-stage, adaptive test strategies where non-inferiority is tested only if the reference shows sufficient efficacy and otherwise δ-superiority of the experimental treatment over placebo is tested. With a properly chosen superiority margin, δ-superiority indirectly shows non-inferiority. We optimize the sample size for several decision rules and find that the natural, data driven test strategy, which tests non-inferiority if the reference’s efficacy test is significant, leads to the smallest overall and placebo sample sizes. We proof that under specific constraints on the sample sizes, this procedure controls the family-wise error rate. All optimal sample sizes are found to meet this constraint. We finally show how to account for a relevant placebo drop-out rate in an efficient way and apply the new test strategy to a real life data set.

KEYWORDS
adaptive test strategy, assay sensitivity, fixed sequence test, “gold standard” non-inferiority trial, multiple testing, optimal sample size

1 | INTRODUCTION

Trial designs with active reference and placebo groups are considered preferable to two-armed non-inferiority designs whenever it is ethical to include the placebo.1-4 The use of a placebo group in non-inferiority trials is particularly useful (and easier to justify) in trials where the placebo effect is expected to be of relevance or it is difficult to identify the population which benefits from the reference treatment. It has therefore been recommended in specific indications like asthma, migraine, depression and panic disorder.5-8 If it is ethically justified, a placebo (P) group is included additionally to the groups of the experimental treatment (E) and reference treatment (R). Frequently, non-inferiority of the new therapy to the reference is sufficient to show, for example if the new therapy has less strong side effects or affects different subgroups of patients. However, non-inferiority can easily be shown if neither the experimental nor the reference therapy has an effect in the study.
The treatment effect observed in a clinical trial depends on several factors, such as the indication under study, the specific trial population, concomitant therapies, and the clinical setting. There are indications in which randomized placebo-controlled trials show a substantial failure rate (e.g., studies in depression). In these indications it is observed, that there is individual variability in the effects of treatments. This leads to a quite large uncertainty about the mean treatment effect to be expected in a specific trial. It may happen, that an established reference treatment proved to be efficacious in previous trials or meta-analyses thereof, but fail in a future trial. In such cases we speak of “unstable reference.”

If the reference is known to be unstable in its effect, then a placebo group is needed to show “assay sensitivity” of the study, that is, that it is planned and conducted in a way that an active treatment, like the reference, will be effective. Koch and Röhmel propose the following hierarchical procedure, which is known as the “gold standard” design for three-armed non-inferiority studies ($\mu_i$ denotes the mean of the effect parameter in group $i \in \{E, R, P\}$, larger means are indicative of more efficacy):

1. Reject $H^S_{EP} : \mu_E - \mu_P \leq 0$, that is, show superiority of $E$ vs $P$.
2. Reject $H^N_{ER} : \mu_E - \mu_R \leq -\delta_N$, that is, show non-inferiority of $E$ vs $R$ with non-inferiority margin $\delta_N$.

In our examples below, we choose the non-inferiority margin $\delta_N$ to be half the historically observed effect of the reference over placebo.

As in Hauschke and Pigeot one may argue that superiority of the experimental treatment over placebo is not a sufficient conclusion if a reference treatment exists. As a more relevant null hypothesis they introduce the hypothesis of $\delta$-superiority:

$$H^\delta_{EP} : \mu_E - \mu_P \leq \delta.$$

One may further argue that the true effect of the reference in the study should decide which of the two hypotheses $H^\delta_{EP}$ or $H^N_{ER}$ is more meaningful to reject after superiority of the new treatment over placebo has been shown. If the reference’s effect is large, then non-inferiority is a strong assertion, which stands for a successful study (i.e., providing evidence about the efficacy of the new treatment). But if the reference is weak, then there is more interest to show $\delta$-superiority of the new treatment, which is a success as well. Moreover, if the non-inferiority margin $\delta_N$ is chosen as the fraction $\rho$ of the historical reference effect (i.e., the historical difference between the reference and placebo is $\delta_N/\rho$), then $\delta = (1 - \rho)\delta_N/\rho$ is the effect of the experimental treatment group that is equivalent to $\delta_N$-non-inferiority. Hence, $\delta$-superiority of the experimental treatment can be taken as (indirect) proof for non-inferiority, also when the reference fails in the given trial. For an illustration see Figure 1. As a consequence, $\delta$-superiority for any larger $\delta$ (e.g., if there is some proposed threshold in the literature) could also be interpreted as indirect proof for non-inferiority.

This argumentation seems to imply a necessity to test the effect of the reference, in order to obtain a reasonable interpretation of the study. However, the sponsor of the study is basically interested in proving the efficacy of the new treatment and does not wish to spend level in showing efficacy of the standard treatment. In this article, we propose an adaptive extension of the Koch-Röhmel design, where the reference effect is not tested but used for a data-driven decision rule (or filter) to decide whether to test for non-inferiority or $\delta$-superiority. As we will see in the following, the above idea of interpretation of study success in dependence of the strength of the reference can be implemented formally via a hierarchical testing procedure that controls the family-wise error rate (FWER), that is, the probability to make at least one erroneous rejection. Sample size calculation and optimization is done with regard to the considered definition of study success. Even though there exists extensive literature and recent research on three-arm non-inferiority trials, we are not aware of a similar suggestion in the literature.

**Figure 1** Illustration of $\delta_N$-non-inferiority of $E$ vs $R$ as a result of $\delta$-superiority of $E$ vs $P$.
The article is organized as follows. In the next section, we introduce the adaptive testing strategy for three-arm trials. The following section considers different rules for deciding upon the strength of the reference, later also called “filter” for the non-inferiority test. It turns out that the intuitive filter requesting a statistical proof of superiority of the reference over placebo yields the smallest sample sizes needed to obtain a certain probability of study success. In further two sections, we adjust our optimal sample size calculation, allowing for uncertainty in the true effect of the reference in the study, and for a relevant dropout rate in the placebo group, respectively. We conclude with a discussion.

2 DESCRIPTION OF THE TEST DESIGN

To demonstrate successfully the efficacy of the new treatment, one would like to proceed as sketched in the left graph of Figure 2. After showing superiority over placebo, the next hypothesis of interest depends on the strength of the reference. If the reference is strong, that is, its effect is larger than that of placebo, then one wants to show non-inferiority of the experimental treatment compared to the reference, which is in the spirit of the gold standard design proposed by Koch and Röhmel. But if the reference is not better than placebo, then non-inferiority to reference is not of interest. Rather, one wants to prove \( \delta \)-superiority to placebo, which indirectly proves the non-inferiority to the historical reference. Note that the decision “\( R > P \)" serves as a filter for further testing and is not part of the test strategy. Several options are possible to define what “\( R > P \)" means (see below).

The intuitive strategy described above does not guarantee that the FWER is controlled. Although only one of the hypotheses \( H_{ER}^N \) or \( H_{EP}^N \) is tested at level \( \alpha \) after rejection of \( H_{EP}^S \), the decision, which of the two hypotheses is tested, is data-driven. Therefore, as a formal implementation of the intuitive idea, we define the following hierarchical test, which is shown in the right graph of Figure 2. This figure uses the notation of graphical test procedures introduced in Bretz et al. and includes the filter as exploratory test to interpret the result. After rejection of \( H_{EP}^S \), we pass the level \( \alpha \) to the non-inferiority hypothesis \( H_{ER}^N \), in concordance with the gold standard design. But in contrast to the proposed continuations in the Koch-Röhmel design, after rejection of the non-inferiority hypothesis, the next hypothesis in our hierarchical sequence to be tested at level \( \alpha \) is the \( \delta \)-superiority \( H_{EP}^\delta \). The filter “\( R > P \)" now serves as interpretation for the success of the study. Success means that we have evidence about the efficacy of the new treatment. This can either be reached if the reference is strong (ie, the filter is satisfied) and non-inferiority of the new treatment over the reference is shown, or if the reference is weak (ie, the filter is not satisfied) and \( \delta \)-superiority of the new treatment is shown. Since \( \delta \) is \((1 - \rho)\) times the historically observed reference effect, this serves as an indirect proof of non-inferiority over the (historical) reference.

For application, the left side of Figure 2 would be easier to explain to practitioners. The hierarchical strategy in the right side of Figure 2 seems unnatural, because one tests first \( E \) vs \( P \), then \( E \) vs \( R \), and then again \( E \) vs \( P \). Assume that we apply the intuitive strategy using the following filter: the reference is declared “strong" if its superiority is shown at significance level \( \alpha \), that is, if

\[
\frac{X_R - X_P}{\sigma \sqrt{n_R^{-1} + n_P^{-1}}} \geq z_{\alpha},
\]

where \( X_R \) and \( X_P \) are the observed group means in the reference and placebo group, \( \sigma \) is the common SD, \( n_R \) and \( n_P \) are the sample sizes in the two groups and \( z_{\alpha} \) is the \((1 - \alpha)\)-quantile of the normal distribution. We will show in Appendix A
that, in a large number of situations, the intuitive strategy with this filter leads to the same decisions concerning success of the study as the formal strategy with the same filter. This seems curious, because the formal strategy requires proof of non-inferiority also in the case, where only $\delta$-superiority of $E$ vs $P$ is of interest, which should reduce the success probability. However, $\delta$-superiority of interest only if the reference is not better than placebo, that is, $X_R \leq X_P$, so that non-inferiority of an effective new treatment is not a high barrier to take.

The equivalence between the two strategies is implied if the following restriction on the sample sizes is satisfied:

$$\left(\sqrt{n_R^{-1} + n_P^{-1}} + \sqrt{n_R^{-1} + n_E^{-1} - \sqrt{n_E^{-1} + n_P^{-1}}\right) z_\alpha \leq (\delta_N + \delta)/\sigma.$$  \hspace{1cm} (2)

A simple check shows that this restriction holds for all optimal sample sizes derived in the remainder of this article. Hence, the new design can be interpreted in an intuitive way, while assuring strong error control. We call this new design single-stage adaptive non-inferiority design for three-arm trials. Note, that while all sample sizes derived in the remainder of this article meet the above condition, a formal proof that any optimal sample sizes for this new study design will meet the condition is beyond the scope of this article and left for further research.

### 3 | FILTER OPTIONS

The filter declaring the reference as “strong” if (1) holds seems to be a natural condition for deciding that we have a strong reference, that is, evidence that “$R > P$” holds in the two graphs of Figure 2. There are some more natural candidates for indicating that the reference has a relevant effect over placebo. For simplicity, we will from now on assume that $\rho = 1/2$, that is, $\delta = \delta_N$ are equal to half of the historical reference effect. We will discuss the following four options (with notations as above):  

- **Filter 1**: $(X_R - X_P)/ \sigma \sqrt{n_R^{-1} + n_P^{-1}} \geq z_\alpha$, that is, superiority as in Equation (1);
- **Filter 2**: $(X_R - X_P - \delta)/ \sigma \sqrt{n_R^{-1} + n_P^{-1}} \geq z_\alpha$, that is, superiority by half the historical reference effect is shown at significance level $\alpha$;
- **Filter 3**: $X_R - X_P \geq 2\delta$, that is, the observed difference of means is at least the historical difference;
- **Filter 4**: $X_R - X_P \geq 3/2\delta$, that is, the observed difference of means is at least $3/4$ of the historical difference.

To evaluate the filter options, we consider the required sample sizes to obtain a probability of 90% for the success of the study, that is, to show either non-inferiority of the new treatment to the reference in the case where the filter is satisfied (denoted by “Power 1” in the following), or to show $\delta$-superiority of the new treatment over placebo in the case where the filter is not satisfied (denoted by “Power 2”). The overall success probability (“Total Power”) is the sum of Power 1 and Power 2.

Schlömer and Brannath have developed an algorithm, which minimizes the total sample size that is needed to obtain a pre-specified power for the original gold standard design, that is, to show non-inferiority of the new treatment compared to the reference. By adjusting this algorithm (see Appendix B), we can calculate optimal sample sizes to reach a desired power for the success of the study. We make the following assumptions for the calculation:

- desired power 90% for the success of the study;
- one-sided FWER $\alpha = 2.5\%$;
- independent normal observations with common SD $\sigma = 0.5$;
- non-inferiority margin $\delta = 0.1$, which corresponds to half of the historical reference effect;
- the effect of the new treatment is equal to the historical reference effect, $\mu_E - \mu_P = 0.2$.

We consider three different scenarios for the effect of the reference over the placebo:

- **Scenario 1**: the reference is as good as observed historically, $\mu_R - \mu_P = 0.2$. 
Table 1 shows the calculated optimal sample sizes for each of the three scenarios and every filter. The most intuitive filter 1, which requires a statistical proof at α-level of superiority of the reference over placebo, leads to the smallest sample sizes in each scenario. In particular, as long as the reference is effective such that non-inferiority suffices for the success of the study, the size of the placebo group is small compared to the other groups, which is ethically favorable. We will therefore in the following consider only the first option for the filter, as given by (1).

Comparing the scenarios, the highest total sample size is needed with all filters in Scenario 1 and the lowest sample size in Scenario 3. In the latter scenario, there is almost no difference between the filters concerning sample size allocation. It is clear that if the reference effect is zero, then success is mainly reached by showing δ-superiority of the experimental treatment over placebo. Hence, the reference sample size is very small with all filters.

When considering Scenario 2, we observe that filters 2-4 all present with almost identical optimal sample sizes. This is due to the fact that in this setup where the reference has only half the historic effect, these filters indicate that “R > P” with almost equal probability (in fact they conclude “R > P” with very small probability). As a consequence, these three filters lead to almost the same testing procedure, needing to show δ-superiority of the experimental treatment over placebo to claim success.

### 4 Accounting for Uncertainty in the Reference Effect

As can be seen from Table 1, sample sizes are quite sensitive to the assumed scenario. This is also illustrated in Table 2, where the contributions of Power 1 (from non-inferiority) and Power 2 (from δ-superiority) to the total power in the three scenarios considered above are indicated: if the reference effect corresponds to the observed historical effect (Scenario 1), then all power is gained by the non-inferiority claim. In the second scenario, superiority of the reference fails to be shown in almost 25% of the cases, therefore some power (14%) must be gained by showing δ-superiority of the new treatment over placebo. If the reference fails completely, then almost all power is gained by the δ-superiority.

Another issue is the question how sensitive the power behaves if the true scenario is actually different from the made assumptions. So far, we presumed that the true scenario is the same as the one assumed for the sample size calculation. Table 3 shows, for the three considered scenarios, that substantial power losses may occur when the assumptions made...
TABLE 2  Different power strategies with optimal sample sizes in the three scenarios

| (Filter 1) | Sc 1 | Sc 2 | Sc 3 |
|------------|------|------|------|
| Filter satisfied | 99.3% | 75.9% | 2.5% |
| Power 1 | 90.0% | 75.6% | 2.2% |
| Power 2 | 0.0% | 14.4% | 87.8% |
| Total power | 90.0% | 90.0% | 90.0% |

TABLE 3  Change of power if true scenario is different from the scenario for which the sample size was calculated

| (Filter 1) | Power in scenario |
|------------|-------------------|
|             | 1     | 2     | 3     |
| Sample size for scenario | 10.0% | 70.8% | 60.1% |
| 2 | 66.6% | 90.0% | 76.3% |
| 3 | 32.3% | 80.0% | 90.0% |

do not hold. This is comprehensible because the optimal sample size allocations between the three groups rely on either of the two different tests to reach the desired power. If the reference effect is weaker or stronger than assumed, then the respective other test is successful, which has not been powered for.

Therefore, we propose an alternative sample size calculation, which is based on Bayesian considerations. Note, that Bayesian arguments are only used for the sample size calculation, the analyses would still be strictly frequentist. Sometimes this is called a “hybrid” Bayesian-frequentist approach.\(^{22,23}\) We consider several possible ratios \(v\) of the reference effect in the study divided by the historical reference effect. Each option for \(v\) is weighted by an assumed probability of its occurrence. More general, denote the density of the probability distribution of \(v\) on \([0, 1]\) with \(f(v)\). Then the weighted success probability is given by

\[
S = \int_{0}^{1} S_v f(v) dv,
\]

where \(S\) is the targeted success probability and \(S_v\) is the success probability if \(v\) is the true ratio. The quantity \(S\) in (3) is often referred to as “probability of success”\(^{23}\) or “assurance.”\(^{24}\)

A very simple application of the above definition is to define a grid of values for \(v\) and assign probabilities for each for each of the values. To illustrate this approach we use a simple example: since we do not expect the reference to fail completely, we expect \(v = 1\) with some probability \(p\) and we assume with probability \((1 - p)/2\) that \(v = 3/4\) and \(v = 1/2\), respectively. The probability of success in (3), which is targeted at 90%, as before, is then given by

\[
S = p S_1 + (1 - p)/2 S_{3/4} + (1 - p)/2 S_{1/2}.
\]

Adjusting the algorithm in Appendix B accordingly (with all other parameters as in the previous section), we obtain new optimal sample sizes for the three groups. Figure 3 (left) illustrates the change of sample sizes for different values of \(p\). In Figure 3 (right), the probability of success is shown in dependence of \(p\) if the true reference effect is \(v = 1, 3/4,\) or 1/2 of the historical effect, respectively.

One sees from Figure 3 that the overall sample size is not monotone in \(p\), the maximum being reached for \(p = 0.75\). The larger \(p\) is, that is, the more likely one expects the reference to be strong, the lower is the optimal sample size for the placebo group. Also, the power shows no monotonicity in \(p\) and likewise not in the true ratio \(v\). The highest power is gained for all \(p\) when the true reference effect is \(v = 3/4\). Without any knowledge of the expected performance of the reference, a probability \(p\) between 0.5 and 0.6 is recommended to be used for the sample size calculation to avoid large power losses in the case of an extremely strong or extremely weak reference.

In practical applications, the choice of the values for \(v\) and \(p\) should be driven by knowledge from prior studies. That is, the values chosen for \(v\) should reflect true ratios observed in prior studies and \(p\) should be driven by the believes as
to how likely the respective values for $v$ will be observed in the planned trial. In practice more complex prior distributions for the parameter $v$ might be used. For example, a uniform distribution on the interval $[0.5, 1]$ could reflect the belief of a moderate reference effect. In that case, the algorithm in Appendix B would need to be adjusted accordingly to derive the optimal sample size: Analogously to what we have shown in Appendix B, $S_v$ can be approximated via the cdf of a 3-dimensional normal distribution. The integral in (3) can then be approximated numerically, for example, by using Monte Carlo methods. Given specific sample size allocations and a probability of success $S$, the required total sample size $N = n_E + n_R + n_P$ can be calculated as the solution to (3). In order to obtain the optimal sample sizes, the required total sample size $N$ is optimized over different sample size allocations, for example, using the method of Byrd et al.

5 | MODIFICATION TO ACCOUNT FOR ETHICAL CONSIDERATIONS

In general, one wants to have a small placebo group in order to offer an effective treatment to most of the patients. The size of the placebo group has also practical relevance, because placebo patients are most likely to leave the study early (even if the study is blinded) so that additional recruitment to the placebo group is necessary. An idea to take the high drop-out rate into account is to include the required extra-recruitment in the optimization procedure. To this end, let $n_P$ be the number of evaluable patients and $w_P n_P$, $w_P > 1$, the required number of recruited patients. Instead of minimizing the sample size $n_P + n_R + n_E$, we target now at $w_P n_P + n_R + n_E$, where $w_P > 1$. When minimizing this quantity to obtain a fixed success probability of the study, the optimal size of the placebo group is reduced, which is in favor of ethical considerations.

We adjusted the algorithm in Appendix B, incorporating both the weighted success probability (4) and a drop-out parameter $w_P$, leaving the other parameters as before. We compare the results from the previous section (where $w_P = 1$) to a setting, where half of the recruited patients in the placebo group drop out ($w_P = 2$). Table 4 compares sample sizes and power results for different values of $p$ and $v$. The probability $p$ for the scenario that the reference is as strong as historically ($v = 1$) is chosen not smaller than $1/2$, because we believe that this is a minimum requirement for the reference in a
clinical trial. The case $p = 1$ excludes the uncertainty about the reference effect and therefore corresponds to Scenario 1 in the section on the filter options.

When assuming $p = 1$ for the sample size calculation, then the sample size of the placebo group is reduced from 159 if $w_P = 1$ to 139 if $w_P = 2$. Of course, we need to initially recruit $2 \times 139 = 278$ patients, but ignoring the drop-out in the optimization procedure one would need to recruit $2 \times 159 = 318$ patients, that is, 40 more, to the placebo group.

On the other hand, the total planned sample size increases from 1244 to 1253. The expected number of recruited patients is $n_E + n_R + 2n_P = 1392$ in the case where this quantity is minimized, compared to 1403 if we do not take the drop-out rate into account. Furthermore, a prize has to be paid in terms of scenario uncertainty. If, for example, the reference in the study has only half the historical effect ($\nu = 0.5$), then the probability of a successful study decreases from 71% if $w_P = 1$ to 66% if $w_P = 2$.

### 6 | EXAMPLE

We next apply the proposed testing procedure to data of a three-arm study on major depressive disorder by Higuchi et al.\textsuperscript{26}

We will first show the example in its original version, as shown in Reference 26. After that, we will apply the proposed procedure to the same data. The primary objective of this double-blinded, randomized, active controlled, parallel-group study was to compare the efficacy and safety of 6-week treatment with duloxetine (E) to those of paroxetine (R) and placebo (P). The primary endpoint of the study was the HAM-D17 change from baseline at 6 weeks and the statistical analysis was planned to test the superiority of duloxetine over placebo and the non-inferiority of duloxetine over paroxetine in hierarchical order. The non-inferiority margin was set as $\delta_N = 2.5$. The observed mean decreases were 10.2 ± 6.1 (mean ± sd) in the duloxetine group, 9.4 ± 6.9 in the paroxetine group and 8.3 ± 5.8 in the placebo group with sample sizes $n_E = 147$, $n_R = 148$, $n_P = 145$. The two-sided 95% confidence interval for the difference in means between duloxetine and placebo is in this case given by (0.53, 3.27), indicating superiority of duloxetine over placebo. The 95% confidence interval for the difference in means between duloxetine and paroxetine can be calculated to be ($-0.69, 2.30$) and excludes $-\delta_N = -2.5$ indicating non-inferiority of duloxetine compared to paroxetine. However, the superiority of paroxetine over placebo could not be established, because the 95% confidence interval for this comparison is given by ($-0.37, 2.57$) and therefore includes 0. Higuchi et al\textsuperscript{26} concluded that non-inferiority of duloxetine compared to paroxetine did not have assay sensitivity.

The trial data were also investigated by Hida and Tango,\textsuperscript{27} who argue that for the proof of assay sensitivity it does not suffice to show that $\mu_R > \mu_E$ but one rather needs to show that the reference is at least $\delta$-superior to placebo ($\mu_R < \mu_E - \delta$). Since, paroxetine could not be shown to be superior to placebo, it is also not considered to be $\delta$-superior. Therefore, Hida and Tango\textsuperscript{27} also conclude the lack of assay sensitivity. It should be noted that in the given setup, the requirement of showing $\delta$-superiority of reference over placebo is very similar to filter 2 introduced above. However, in the approach of Hida and Tango,\textsuperscript{27} the testing procedure stops if $\delta$-superiority cannot be established, while in our proposed approach, we
would still have the opportunity to obtain a successful study by showing \( \delta \)-superiority of the experimental treatment over placebo. Following the approach outlined in the introduction (cf. Figure 1) \( \delta \) is chosen in a way that \( \delta \)-superiority of \( E \) over \( P \) is equivalent to \( \delta_N \)-non-inferiority of \( E \) to the historical reference effect. In their sample size calculation Higuchi et al\(^ {26} \) assume a historical effect of \( R \) over \( P \) of 5. Therefore, in our application we chose \( \delta = \delta_N = 2.5 \).

### 6.1 Application of the proposed method

We now apply the proposed testing procedure to the data mentioned above. As stated above, duloxetine can be shown to be superior to placebo (i.e., \( H_{EP}^R \) is rejected). However, using filter 1, “\( R > P \)” cannot be concluded, which is why as a next step we test for \( \delta \)-superiority of duloxetine over placebo (\( H_{EP}^\delta \)). Since the 95% confidence interval for this comparison includes \( \delta = 2.5 \), \( H_{EP}^\delta \) cannot be rejected. Therefore, for the given data, using the newly proposed methods leads to the same conclusion as the methods applied originally. Note that we applied the intuitive testing approach, instead of the formal definition. In the case of this study both approaches are equivalent (because the sample sizes fulfill the special condition derived in Appendix A).

For illustrative purposes we assume that the observed mean decrease in HAM-D17 at week 6 in the duloxetine group was 12.2 ± 6.1 (mean ± SD). We re-calculate the confidence interval for the comparison of the duloxetine and placebo groups as (2.53, 5.27). Under this assumption, duloxetine is of course still superior to placebo. Furthermore, using filter 1 “\( R > P \)” cannot be concluded, which is why as a next step we test for \( \delta \)-superiority of duloxetine over placebo, which in this case can be concluded since the aforementioned confidence interval excludes \( \delta = 2.5 \). Hence, the study would be successful in the sense that, the experimental treatment could be shown to be sufficiently effective. Note that the standard approach to the testing of non-inferiority as well as the approach by Hida and Tango\(^ {27} \) still conclude the lack of assay sensitivity and stop there, without considering the fact that the experimental treatment is sufficiently effective to claim success even when the reference is weak.

Note, that the confidence intervals reported in this section, while appropriate for deriving the test decision in the hierarchical test, are no simultaneous confidence intervals, and therefore do not have simultaneous coverage probability. If one is interested in informative simultaneous confidence intervals, the procedure of Schmidt and Brannath\(^ {28} \) could be implemented. In that case appropriate splitting weights would have to be defined, which is beyond the scope of this article. Future research on this topic might involve defining these weights in dependence on the filter chosen in our procedure.

### 6.2 Sample size

For the calculation of the optimal sample size, we use the assumptions proposed by Hida and Tango\(^ {27} \): (i) duloxetine and paroxetine have the same effect size (\( \mu_E = \mu_R = 10 \)), (ii) the placebo has half the effect (\( \mu_P = 5 \)), (iii) the three arms have a common SD \( \sigma = 6.5 \) and (iv) the same value as in the above-stated study for the non-inferiority margin is chosen: \( \delta = 2.5 \). Under these assumptions, the optimal sample size to achieve a power of 80% (which is the same as assumed in Hida and Tango\(^ {27} \)) for our proposed method at the one-sided significance level 0.025 is: \( n_E = 110 \), \( n_R = 114 \), \( n_P = 39 \), resulting in an overall sample size of \( N = 263 \) with filter 1 and \( n_E = 130 \), \( n_R = 131 \), \( n_P = 101 \) resulting in an overall sample size of \( N = 362 \) with filter 2. These sample sizes are considerably smaller than those calculated by Hida and Tango\(^ {27} \) who need \( n_E = n_R = 151 \), \( n_P = 121 \) (\( N = 423 \)) under the same assumptions. Furthermore, both sample sizes for the new approach fulfill the condition (2), leading to equivalence between the intuitive design and the formal design for the new testing procedure.

For illustrative purposes, we also apply the sample size determination procedure accounting for uncertainty in the reference effect introduced in Section 4. We assume the same values for \( \mu_E = 10 \), \( \mu_P = 5 \), \( \sigma = 6.5 \), and \( \delta = 2.5 \) as before. Suppose that from previous studies we know that only in 50% of the trials we observe the full historical effect of the reference. This means \( v = 1 \) (resp. \( \mu_R = 10 \)) with probability \( 1/2 \). In the other 50% of trials the drug under investigation is known to have a weaker effect of \( \mu_R = 7 \) (i.e., \( v = 0.4 \)). Under these assumptions, the optimal sample size to achieve a power of 80% at the one-sided significance level 0.025 is: \( n_E = 109 \), \( n_R = 94 \), \( n_P = 87 \), resulting in an overall sample size of \( N = 290 \) with filter 1. The uncertainty about the reference effect leads to these sample sizes being larger than those derived under the assumption of a stable “full” reference effect of \( \mu_R = 10 \). However, these sample sizes are still
considerably smaller than those calculated by Hida and Tango and also fulfill the condition (2), leading to equivalence between the intuitive design and the formal design for the new testing procedure.

7 SUMMARY AND DISCUSSION

The gold standard design is applied in indications with an unstable reference. Proving efficacy of the reference is not its main goal and only necessary if success of the new treatment is shown via non-inferiority compared to reference. This idea is made explicit with the adaptive testing strategy introduced here. It provides the possibility to assure success also with a weak reference by showing δ-superiority of the new treatment over placebo.

A flexibility is proposed also with respect to the uncertainty of the reference performance by underlying a scenario mixture in the sample size calculation. We propose to assume a strong or partly strong reference with probability around one half, respectively, yielding high success probabilities for any true reference effect. Furthermore, an adjustment of sample size is considered for the case of a high drop-out rate in the placebo group.

The new design can easily be communicated to practitioners by an intuitive graphic. Here, it has to be taken care that the sample sizes satisfy a certain condition (see (2) and Appendix A), which guarantees equivalence of the intuitive picture to a more formal one and thus control of the family-wise error rate. This was never a problem in all settings considered here. Further investigation might be done to make this formal condition more comprehensible, to find a formal proof that all optimal sample sizes meet the condition or maybe to adapt the filter in a way that no such condition is necessary.

The idea of testing non-inferiority and δ-superiority may alternatively be implemented within a Bayesian framework. Here we could use the joint posterior distribution of the three effect parameters to calculate for example, the posterior probability that the experimental treatment is non-inferior to the reference and the reference is superior to placebo and the posterior probability that the experimental treatment is δ-superior to placebo. The treatment could then be claimed useful if at least one these two posterior probabilities are sufficiently high. However, since confirmatory claims often rely on frequentist methods, we focused on the frequentist approach.

A further topic of future research might be deriving informative simultaneous confidence intervals in the framework defined in this article.

ACKNOWLEDGEMENT

We thank Anna Schritz for her help in the development of the R programs which were used for this manuscript. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

The authors disclosed receipt from the following financial support for the research, authorship, and/or authorship of this article: This work was supported by the German Research Foundation (DFG) under Grant BR 3737/1-1.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Werner Brannath https://orcid.org/0000-0002-8622-3904
Martin Scharpenberg https://orcid.org/0000-0003-1584-0056
Sylvia Schmidt https://orcid.org/0000-0001-8090-8083

REFERENCES

1. CPMP. ICH Topic E 10. Choice of Control Group in Clinical Trials. CPMP/ICH/364/96; 2001. European Medicines Agency, London.
2. CPMP. Guideline on the choice of the non-inferiority margin. EMEA/CPMP/EWP/2158/99; 2005. European Medicines Agency, London.
3. EMA. Reflection paper on the need for active control in therapeutic areas where use of placebo is deemed ethical and one or more established medicines are available. Draft, EMA/759784/2010; 2010. European Medicines Agency, London.
4. EMEA. EMEA/CPMP position statement on the use of placebo in clinical trials with regard to the revised declaration of Helsinki. EMEA/17424/01; 2001; The European Agency for the Evaluation of Medical Products, London.
5. CPMP. Note on guidance on the clinical investigation of medicinal products in the treatment of asthma. CPMP/EWP/2922/01; 2002. European Medicines Agency, London.
6. CPMP. Note on guidance on the clinical investigation of medicinal products in the treatment of panic disorder. CPMP/EWP/4280/02; 2005. European Medicines Agency, London.
7. CPMP. Note on guidance on the clinical investigation of medicinal products in the treatment of depression. CPMP/EWP/788/01; 2007; European Medicines Agency, London.
8. Kaiser T, Volkmann C, Karyotaki E, Cuijpers P, Brakemeier EL. Heterogeneity of treatment effects in trials on psychotherapy of depression. APA PsycArticles; 2022; Advance Online Publication. 10.1037/cps0000079
9. Koch A, Röhmel J. Hypothesis testing in the “gold standard” design for proving the efficacy of an experimental treatment relative to placebo and a reference. J Biopharm Stat. 2004;14(2):315-325. doi: 10.1081/BIP‐120037182
10. Hauschke D, Pigeot I. Establishing efficacy of a new experimental treatment in the 'gold standard' design. Biom J. 2005;47(6):782-786. doi:10.1002/bimj.200510169
11. Mütze T, Konietschke F, Munk A, Friede T. A studentized permutation test for three-arm trials in the ‘gold standard’ design. Stat Med. 2017;36:883-898.
12. Mütze T, Friede T. Blinded sample size re-estimation in three-arm trials with 'gold standard' design. Stat Med. 2017;36:3636-3653. doi:10.1002/sim.7356
13. Wu Y, Li Y, Hou Y, Li K, Zhou X. Study duration for three-arm non-inferiority survival trials designed for accrual by cohorts. Stat Methods Med Res. 2018;27:507-520. doi:10.1177/0962280216633908
14. Zhong J, Wen MJ, Kwong KS, Cheung SH. Testing of non-inferiority and superiority for three-arm clinical studies with multiple experimental treatments. Stat Methods Med Res. 2018;27:1751-1765. doi:10.1177/096228021668913
15. Xu W, Hu F, Cheung SH. Adaptive designs for non-inferiority trials with multiple experimental treatments. Stat Methods Med Res. 2018;27:3255-3270. doi:10.1177/0962280217695579
16. Lu H, Jin H, Zeng W. A more efficient three-arm non-inferiority test based on pooled estimators of the homogeneous variance. Stat Methods Med Res. 2018;27:2437-2446. doi:10.1177/0962280216681036
17. Homma G, Diamon T. Sequential parallel comparison design for ”gold standard” noninferiority trials with a prespecified margin. Biometr J. 2019;61:1493-1506.
18. Homma G, Diamon T. Sample size calculation for “Gold-standard” noninferiority trials with fixed margins and negative binomial endpoints. Stat Biopharm Res. 2021;13:435-447.
19. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. Stat Med. 2009;28(4):586-604. doi:10.1002/sim.3495
20. Schläfer P, Brannath W. Group sequential designs for three-arm 'gold standard' non-inferiority trials with fixed margin. Stat Med. 2013;32(28):4875-4889. doi:10.1002/sim.5950
21. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-care Evaluation. Hoboken: Wiley; 2004.
22. Kunmann K, Grayling MJ, Lee KM, Robertson DS, Rufibach K, Wason JMS. A review of Bayesian perspectives on sample size derivation for confirmatory trials. Am Stat. 2021;75(4):424-432. doi:10.1080/00031305.2021.1901782
23. Stallard N, Posch M, Friede T, Koenig F, Brannath W. Optimal choice of the number of treatments to be included in a clinical trial. Stat Med. 2009;28:1321-1338. doi:10.1002/sim.3551
24. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. SIAM J Sci Comput. 1995;16:1190-1208. doi:10.1137/0916069
25. Higuchi T, Murasaki M, Kamijima K. Clinical evaluation of duloxetine in the treatment of major depressive disorder-placebo- and paroxetine-controlled double-blind comparative study. Jpn J Clin Psychopharmacol. 2009;12:1613-1634.
26. Schmidt S, Brannath W. Informative simultaneous confidence intervals in hierarchical testing. Methods Inf Med. 2014;53:278-283.

How to cite this article: Brannath W, Scharpenberg M, Schmidt S. Single-stage, three-arm, adaptive test strategies for non-inferiority trials with an unstable reference. Statistics in Medicine. 2022;41(25):5033-5045. doi: 10.1002/sim.9552
APPENDIX A. COMPARISON OF THE INTUITIVE AND THE FORMAL DESIGN

Our discussion restricts to the preferred Filter 1, other filters can be handled similarly. We use in this section the following notations for the relevant test statistics:

\[
T_{EP}^\delta = \frac{X_E - X_P - \delta}{\sigma \sqrt{n_E^{-1} + n_P^{-1}}}, \quad T_{RP}^\delta = \frac{X_R - X_P}{\sigma \sqrt{n_R^{-1} + n_P^{-1}}}, \quad T_{ER}^N = \frac{X_E - X_R + \delta_N}{\sigma \sqrt{n_E^{-1} + n_R^{-1}}}
\]

These test statistics are not independent, in fact,

\[
T_{KP}^S = \sqrt{n_E^{-1} + n_P^{-1}} T_{EP}^\delta - \sqrt{n_R^{-1} + n_P^{-1}} T_{ER}^N + (\delta_N + \delta)/\sigma
\]

(A1)

The only difference between the formal and the intuitive approach is that the latter interprets observations as leading to a success of the study, where the filter is not satisfied and \(\delta\)-superiority of \(E\) vs. \(P\) can be shown, whereas formally it is additionally necessary that \(E\) is non-inferior to \(R\) because of the hierarchical structure, which guarantees error control. Hence, different decisions with the two strategies occur if and only if

\[
T_{RP}^S < z_\alpha \quad \text{and} \quad T_{EP}^\delta \geq z_\alpha \quad \text{and} \quad T_{ER}^N < z_\alpha.
\]  

(A2)

By (A1), we have

\[
T_{RP}^S < z_\alpha \quad \Leftrightarrow \quad T_{ER}^N > \sqrt{n_E^{-1} + n_P^{-1}} T_{EP}^\delta - \sqrt{n_R^{-1} + n_P^{-1}} z_\alpha + (\delta_N + \delta)/\sigma
\]

\[= : B.\]

Now, the condition \(T_{ER}^N < z_\alpha\) implies \(B < z_\alpha\), or equivalently,

\[
T_{EP}^\delta < \left( \sqrt{n_E^{-1} + n_R^{-1}} + \sqrt{n_R^{-1} + n_P^{-1}} \right) z_\alpha - (\delta_N + \delta)/\sigma
\]

\[= : C.\]

Furthermore, \(T_{EP}^\delta \geq z_\alpha\) necessitates \(z_\alpha < C\). As a consequence, situation (A2) cannot occur when \(z_\alpha \geq C\), and then the naive and formal procedures coincide. One easily verifies that \(z_\alpha \geq C\) is equivalent to

\[
\left( \sqrt{n_E^{-1} + n_R^{-1}} + \sqrt{n_R^{-1} + n_P^{-1}} - \sqrt{n_E^{-1} + n_P^{-1}} \right) z_\alpha \leq (\delta_N + \delta)/\sigma.
\]  

(A3)

This condition depends only on the sample sizes of the three groups. It is generally satisfied in interesting situations, where basically the reference group is not too small. A closer look shows that the condition is satisfied for all optimal sample sizes derived in this article. Furthermore, we have determined that the optimal sample sizes when applying the intuitive design are exactly the same as for the formal design in all situations discussed here. Hence, it is very often not necessary to distinguish between the two designs so that the new adaptive design will be easier to communicate to practitioners.
APPENDIX B. ALGORITHM FOR DETERMINATION OF OPTIMAL SAMPLE SIZES

In this section we derive the algorithm for the determination of optimal sample sizes in the proposed new study design. The algorithm is based on Schlömer and Brannath$^{21}$ and adjusted to the new design. Let $c_R = n_R/n_E$ and $c_P = n_P/n_E$. For large sample sizes $n_E, n_R,$ and $n_P$ the vector of test statistics $(T_{EP}, T_{RP}^S, T_{ER}^N)'$, where

$$T_{EP} = \frac{X_E - X_P}{\sqrt{\frac{c_P}{1 + c_P} n_E} \cdot \sqrt{\frac{c_P c_R}{c_P + c_R} n_E} \cdot \frac{\mu_E - \mu_P}{\sigma} \sqrt{\frac{c_P}{1 + c_P} n_E} \cdot \frac{\mu_E - \mu_R + \delta_N}{\sigma} \sqrt{\frac{c_R}{1 + c_R} n_E}}$$

and $T_{RP}^S$ and $T_{ER}^N$ as defined in Appendix A is asymptotically multivariate normally distributed with mean vector

$$\mu = \left( \frac{\mu_E - \mu_P}{\sigma} \sqrt{\frac{c_P}{1 + c_P} n_E}, \frac{\mu_R - \mu_P}{\sigma} \sqrt{\frac{c_P c_R}{c_P + c_R} n_E}, \frac{\mu_E - \mu_R + \delta_N}{\sigma} \sqrt{\frac{c_R}{1 + c_R} n_E} \right),$$

and covariance matrix $\Sigma$. Here,

$$\Sigma = \begin{pmatrix} 1 & \rho_{EPFR} & \rho_{EPER} \\ \rho_{EPFR} & 1 & \rho_{RPER} \\ \rho_{EPER} & \rho_{RPER} & 1 \end{pmatrix}$$

with $\rho_{EPFR} = \sqrt{\frac{c_P}{1 + c_P} (1 + c_R)}$, $\rho_{EPER} = \sqrt{\frac{c_P c_R}{1 + c_P} (1 + c_R)}$, and $\rho_{RPER} = \sqrt{\frac{c_P}{1 + c_P} (1 + c_R)}$. The probability of success can thus be approximated by

$$S \approx P \left( \{ T_{EP} \geq z_\alpha \} \cap \{ T_{ER}^N \geq z_\alpha \} \cap \{ \text{Filter} : R > P \} \right) + P \left( \{ T_{EP} \geq z_\alpha + \frac{\delta}{\sigma} \sqrt{\frac{c_P}{1 + c_P} n_E} \} \cap \{ T_{ER}^N \geq z_\alpha \} \cap \{ \text{Filter} : R \leq P \} \right).$$

Recall that, the first term is named “Power 1” and the second term is named “Power 2” in Section 3. When using Filter 1, the above equation results to:

$$S \approx P \left( \{ T_{EP} \geq z_\alpha \} \cap \{ T_{ER}^N \geq z_\alpha \} \cap \{ T_{RP} \geq z_\alpha \} \right) + P \left( \{ T_{EP} \geq z_\alpha + \frac{\delta}{\sigma} \sqrt{\frac{c_P}{1 + c_P} n_E} \} \cap \{ T_{ER}^N \geq z_\alpha \} \cap \{ T_{RP} < z_\alpha \} \right). \quad (B1)$$

For the other filters discussed in this article, similar representations can be found. We calculate the overall probability of success $S$ by means of the cumulative distribution function of the multivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma$. Given specific allocation ratios $c_R, c_P$ and a probability of success $S$, the required total sample size $N = n_E + n_R + n_P = (1 + c_R + c_P)n_E$ can be calculated as the solution to (B1). In order to obtain the optimal sample sizes, the required total sample size $N$ is optimized over $c_R$ and $c_P$ using the method of Byrd et al.$^{25}$