

Optimization of adaptive designs with respect to a performance score

Carolin Herrmann | Meinhard Kieser | Geraldine Rauch | Maximilian Pilz

1 Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Berlin, Germany
2 Institute of Medical Biometry, University Hospital Heidelberg, Heidelberg, Germany

Correspondence
Carolin Herrmann, Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, Berlin 10117, Germany. Email: carolin.herrmann@charite.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: KI 708/4-1, RA 2347/4-1, KI 708/4-2, RA 2347/4-2

Abstract
Adaptive designs are an increasingly popular method for the adaptation of design aspects in clinical trials, such as the sample size. Scoring different adaptive designs helps to make an appropriate choice among the numerous existing adaptive design methods. Several scores have been proposed to evaluate adaptive designs. Moreover, it is possible to determine optimal two-stage adaptive designs with respect to a customized objective score by solving a constrained optimization problem. In this paper, we use the conditional performance score by Herrmann et al. (2020) as the optimization criterion to derive optimal adaptive two-stage designs. We investigate variations of the original performance score, for example, by assigning different weights to the score components and by incorporating prior assumptions on the effect size. We further investigate a setting where the optimization framework is extended by a global power constraint, and additional optimization of the critical value function next to the stage-two sample size is performed. Those evaluations with respect to the sample size curves and the resulting design’s performance can contribute to facilitate the score’s usage in practice.

Keywords
adaptive design, clinical trial, optimal design, performance score, sample size calculation

1 | INTRODUCTION

The determination of the sample size in a clinical trial is a crucial aspect. While too small sample sizes lead to underpowered trials and are thus ethically questionable, too large sample sizes can put participants unnecessarily at risk. Chow et al. (2018) state that many sample size calculations presented in anesthesiology journals “often assumed effect sizes larger than those actually observed in a study.” Hence, there is an inherent risk of falsly specifying the sample size in the planning phase of a trial. Updating the sample size during an ongoing trial therefore seems an attractive option. Adaptive designs provide a statistical solution for dealing with data-driven midtrial modifications without undermining the type I error rate. One possible modification is the adaptation of the sample size. More precisely, in an interim analysis, the trial...
is either stopped early or the sample size can be adapted. Within this paper, we focus on two-stage adaptive designs with a single unblinded interim analysis.

The fundamental theory of adaptive designs is well developed, and overviews can be found, for example, in Bauer et al. (2016) or Wassmer and Brannath (2016). Moreover, there also exists a variety of sample size recalculation rules. Many sample size recalculation rules rely on conditional power (Hsiao et al., 2019; Mehta & Pocock, 2011; Proschan & Hun Berger, 1995). The conditional power describes the probability of correctly rejecting the null hypothesis at the end of the trial, given the data observed at interim. However, sample size calculation based on conditional power arguments is also controversially discussed in the literature (Bartroff & Lai, 2008; Bauer & König, 2006; Levin et al., 2013). Other sample size recalculation approaches include further criteria, for example, economic ones (Mehta & Patel, 2006). Various publications on the optimization of adaptive designs exist. Brannath and Bauer (2004) presented optimal conditional error functions under the sample size recalculation rule of Proschan and Hunsberger (1995). Jennison and Turnbull (2006) optimized adaptive designs including the conditional value function assuming different prior distributions on the effect size. The efficiency of optimal group-sequential and optimal adaptive designs was, among others, compared by Jennison and Turnbull (2006) and Lokhnygina and Tsiatis (2008). Jennison and Turnbull (2015) considered sample size recalculation as an optimization problem. Their approach was extended by Pilz et al. (2019). Furthermore, the R-package adoptr (Kunzmann et al., 2021) allows the computation of adaptive designs as solutions to optimization problems with respect to specific objective criteria. The underlying idea is that a two-stage adaptive design can be characterized by the stage-one and stage-two sample size as well as futility and efficacy boundaries. Those parameters can then be determined optimally with respect to a customized objective score under additional constraints, like a minimal power value. The numerical approach is described in Kunzmann et al. (2021). A broad variety of examples for this approach is presented in Pilz et al. (2021).

While the software solution to the above-described optimization approach with respect to specific objective criteria presents a highly flexible sample size solution, it is at the same time enormously dependent on the predefined scoring criteria judging optimality. Hence, the discussion of appropriate scoring criteria is inevitable. Furthermore, despite the variety of sample size recalculation approaches in the literature, there exist no clear guidelines on the scoring and comparison of such sample size recalculation rules. Performance evaluation in adaptive designs comes with the challenge of different evaluation perspectives. Several performance scores were suggested to evaluate sample size recalculation from the global perspective (Fang et al., 2018; Liu et al., 2008; Wu & Cui, 2012). However, the conditional perspective seems also appealing when comparing different stage-two sample size functions. Here, the term “conditional” refers to those cases where the interim test statistic suggests a trial continuation. Herrmann et al. (2020) suggested a performance score evaluating sample size recalculation rules from this conditional perspective. Next to the evaluation of conditional power and second stage sample size with respect to location, they also included measures for the variations thereof. To deeper analyze this score, it seems natural to look at the methodological aspect of deriving stage-two sample size curves by using the constrained optimization approach underlying the R-package adoptr with the conditional performance score by Herrmann et al. (2020) as a scoring criterion. Moreover, in previous work, the performance score was only applied to point prior distributions and is in the following also applied to continuous distributions. At the same time, the individually feasible optimal score value determined by the optimization approach can also be seen as the best conditional performance score value for a specific adaptive design setting under a given prior. This can serve as help when comparing different established sample size recalculation rules with respect to the conditional performance score and how good they perform compared to the best possible performance value in that setting.

In the following, we elaborate on these ideas by first describing the underlying setting and conditional performance score with a point prior under the alternative. We then evaluate the influence of other prior distributions under the alternative hypothesis. Moreover, we derive optimal designs with respect to the conditional performance score with a different weighting of the score components, additional unconditional constraints in the optimization problem as well as a variation of the stage-two critical value function. Finally, we summarize aspects to consider when deriving optimal adaptive designs with respect to the conditional performance score.

2 NOTATION AND INITIAL SETTING

In this paper, we assume two groups T and C with normally distributed outcomes $X_T^I \sim \mathcal{N}(\mu_T, \sigma^2)$, $X_C^I \sim \mathcal{N}(\mu_C, \sigma^2)$ and common, known variance $\sigma^2$. The two groups are compared by a standard $Z$-test. The testing problem is defined
Thus, without loss of generality, higher outcomes are seen as favorable. The adaptive design is defined as follows. After the primary endpoint of \( n_1 \) patients per group has been observed, the test statistic \( Z_1 \) is computed. The null hypothesis is rejected, and the trial is stopped for efficacy if \( Z_1 > c_e \), where \( c_e \) defines the critical value at interim. If \( Z_1 < c_f \), the trial is stopped early for futility without rejection of the null hypothesis. If \( c_f \leq Z_1 < c_e \), the trial enters the second stage. In this case, the stage-two sample size \( n_2 \) can be chosen in dependence of \( Z_1 \) and is, therefore, a function \( n_2(Z_1) \). At the final analysis, the null hypothesis is rejected if \( Z_2 > c_2(Z_1) \) where \( Z_2 \) denotes the test statistic computed from the data of the \( n_2(Z_1) \) patients and \( c_2(Z_1) \) is a critical value that is also chosen in dependence of \( Z_1 \). In total, an adaptive two-stage design can thus be defined as a five-tuple \( D = (n_1, c_f, c_e, n_2(\cdot), c_2(\cdot)) \).

There exist various approaches to the concrete choice of the design parameters. For instance, the early stopping boundaries can be chosen according to the rules by Pocock (1977) or O’Brien and Fleming (1979). The stage-two sample size may be chosen constant (leading to a classical group-sequential design), or its choice may be driven by conditional power considerations, for example, Proschan and Hunsberger (1995). For the critical value function \( c_2(\cdot) \), combination tests as proposed by Bauer and Köhne (1994) or Lehmacher and Wassmer (1999) are frequently used.

For illustrative purposes, we introduce the following initial setting. For now, the standardized expected effect size under the alternative, which is used for power calculations, is assumed to equal \( \delta := (\mu_T - \mu_C)/\sigma = 0.4 \). The maximum sample size is set to \( n_{\text{max}} = 500 \). The trial starts with a per-group sample size of \( n_1 = 50 \) patients. If a negative effect is observed in the interim, the trial is stopped for futility, that is, \( c_f = 0 \). The early efficacy stop is chosen according to Pocock (1977) as \( c_e = 2.178 \) under a global one-sided significance level \( \alpha = 2.5\% \). The stage-two sample size is constant with \( n_2 = 61 \), and the critical value function \( c_2(\cdot) \) follows the inverse normal rule with equal weights \( 1/\sqrt{2} \) and is therefore defined as

\[
c_2(z_1) = \frac{2.178 - z_1/\sqrt{2}}{1/\sqrt{2}}.
\]  

These choices ensure that the maximum type I error rate is protected at a one-sided global level \( \alpha = 2.5\% \), and that a power of \( 1 - \beta = 80\% \) is achieved at the point alternative \( \delta = 0.4 \). In the following, we denote this design by \( D_0 \). It is graphically illustrated by the solid line in Figure 1.

Herrmann et al. (2020) proposed a new conditional performance score that rates the expected sample size, its variability, the conditional power, and its variability of an adaptive design \( D \) at a prespecified effect size \( \delta \). Note that the word “conditional” means “conditional on entering the second stage” and not on a particular stage one result. By weighting four score components, the score is defined by

\[
S(D, \delta) := \omega_{e,p} s_{e,p}(D, \delta) + \omega_{v,n} s_{v,n}(D, \delta) + \omega_{c,p} s_{c,p}(D, \delta) + \omega_{v,c} s_{v,c}(D, \delta).
\]
The concrete definition of the score is provided in Appendix A. The score components are constructed such that higher score values refer to a better performing design. By requiring $\omega_n \geq 0$, $\omega_1 \geq 0$, and $\omega_{cp} \geq 0$, it is ensured that $S(D, \delta) \in [0, 1]$ for each two-stage design $D$ and effect size $\delta$. In the default setting of the conditional performance score, it holds that $\omega_n = \omega_1 = \omega_{cp} = 0.25$. We denote the score in this case by $S_0$. Using these score weights implies that the above-defined design is rated as $S_0(D_0, 0.4) = 0.75273$.

### 3 Optimization of the Performance Score

Once a performance measure for an adaptive design is chosen, one may strive for finding the design that is optimizing this criterion. In this paper, we present and evaluate the optimization with respect to the conditional score by Herrmann et al. (2020) denoted by $S(D, \delta)$ as the performance measure. Since higher score values indicate better performance, the score is to be maximized. Note that we assume the first-stage parameters $n_1, c_f$, and $c_e$ to be fixed in compliance with the conditional character of the score. Furthermore, $c_2(\cdot)$ is fixed by the inverse normal combination test following the setting of Herrmann et al. (2020). Consequently, the tuning parameter to optimize the score is the stage-two sample size function $n_2(\cdot)$. The optimization problem to solve is then:

$$\max_{n_2(\cdot)} S(D, \delta).$$

This problem can be solved using numerical techniques implemented in the R-package adoptr (Kunzmann et al., 2021).

The resulting design $D_1$ for $S = S_0$ with $\delta = 0.4$ is depicted in Figure 1 (dashed line). The sample size function is no longer constant but unimodal. This implies that the observed conditional power curve is flattened and, thus, has a smaller range of possible values.

The resulting optimal design is remarkable for different reasons. It achieves a score value of $S_0(D_1, 0.4) = 0.76835$. Therefore, the optimal design shows only a slight improvement compared with the initially chosen group-sequential design, which achieved a score value of $S_0(D_0, 0.4) = 0.75273$. In detail, the initial group-sequential design achieves 98.0% of the optimally achievable score value. Furthermore, the optimal score value is considerably below 1. This implies that the maximum score value of 1 is clearly not achievable in this setting and that, therefore, all other designs’ ratings have to be seen under this perspective. Thus, a score of 0.75 indicates a very high performance here since it is close to the maximally achievable value.

This optimization approach can now be used to investigate different variations of the score $S$. Instead of evaluating the score $S$ at a single effect size $\delta$, one may think about using a prior distribution $\delta \sim \pi(\delta)$ that is not a simple point prior at $\delta = 0.4$ to prevent overfitting to this particular effect size. Furthermore, different weightings of the score components are imaginable. These ideas are presented in more detail in the following.

### 4 Robustification by Prior Distributions

So far, we have assumed a point prior at $\delta = 0.4$. Using a point prior may to some extent scrutinize the necessity of an adaptive design, since there either exists a rather precise assumption about the effect size or a minimally clinically relevant effect size is possible to define. Consequently, it seems natural to use a continuous prior $\delta \sim \pi(\delta)$ to address the lack of knowledge about the effect size more directly. In this case, the conditional score $S$ can be integrated over the prior and multiple effect sizes are taken into account when a design is rated.

Furthermore, a continuous prior distribution allows the combination of prior knowledge (i.e., the prior) and interim results (the test statistic $Z_1$) by applying Bayes’ formula. Instead of having to choose between the assumed conditional power (conditional power under the planning point prior) and the observed conditional power (conditional power under the point effect observed in the interim), the predictive power (Spiegelhalter et al., 1986) can be used. This means that the conditional power subscores are from now on based on

$$PP(z_1, \pi) = P_{\pi}[\text{reject } H_0 \mid Z_1 = z_1] = \int P_{\delta}[\text{reject } H_0 \mid Z_1 = z_1] \pi(\delta \mid z_1) \, d\delta,$$
where $\pi_{\text{post}}$ is the posterior density

$$\pi_{\text{post}}(\delta | z_1) \propto f_{\delta}(z_1)\pi(\delta)$$

(6)

and $f_{\delta}$ denotes the probability density of $Z_1$ under the effect size $\delta$.

Besides the point prior $\pi_{\text{point}}(\delta) = 1(\delta = 0.4)$, we consider two continuous prior distributions in the following. At first, we assume a truncated normal prior on the interval $[0.1, 0.7]$ with mean value 0.4 and standard deviation 0.2. Second, we investigate a uniform prior on the interval $[0.1, 0.7]$. Therefore, we define

$$\pi_{\text{normal}}(\delta) \sim \mathcal{N}_{[0.1,0.7]}(0.4, 0.2) \quad \text{and} \quad \pi_{\text{uniform}}(\delta) = \mathcal{U}([0.1, 0.7]).$$

(7)

In Figure 2, the three prior distributions are plotted (right panel). The score was optimized under these three prior distributions, and the resulting sample size functions are illustrated in Figure 2 (left panel).

The sample size functions vary less under the two continuous prior distributions. In the case of a uniform prior, higher sample sizes are observed. This may be caused by the fact that for a heavy-tailed continuous prior, larger sample sizes are needed to achieve the same conditional power level as for a point prior (Dallow & Fina, 2011).

Interestingly, the overall score value does hardly differ between the three designs. In the middle panel of Figure 2, the score value is depicted for each design. The score is computed assuming a point prior to the respective value of $\delta$. Due to the different shapes of the sample size of the three designs, having a look at the four subscores of $S_0$ may lead to informative results. Table 1 shows that the subscores of the respective optimal designs behave quite differently.

We observe that all three designs show a very large expected sample size component $s_{e,n}$. The value of the expected conditional power component $s_{e,cp}$ becomes quite low under the two designs with a continuous prior. The variation components are smallest under the uniform prior. In particular, the variation component of the sample size, $s_{v,n}$, is remarkably decreased under the uniform prior compared with the point and the normal prior.

The fairly different sample sizes for the uniform prior in comparison to the point and normal prior reflect accounting for more insecurity about the effect size. To evaluate it in more detail, we also included a truncation to the interval

**TABLE 1** Score components of optimal designs under different prior distributions

| Prior      | Design | $s_{e,n}$ | $s_{v,n}$ | $s_{e,cp}$ | $s_{v,cp}$ | Score       |
|------------|--------|-----------|-----------|------------|------------|-------------|
| Point      | $D_1$  | 0.99987   | 0.86180   | 0.52097    | 0.68936    | 0.76835     |
| Normal     | $D_2$  | 0.99976   | 0.83301   | 0.34922    | 0.75846    | 0.73514     |
| Uniform    | $D_3$  | 0.99975   | 0.70376   | 0.38756    | 0.71607    | 0.70178     |

- $D_1$: point prior
- $D_2$: normal prior
- $D_3$: uniform prior

**FIGURE 2** Optimal designs under different prior distributions
We can see in Figure 3 that the truncation to that smaller interval leads to much more similar sample size curves and that the conditional score is monotonically increasing in the considered interval.

Especially under the uniform prior, there is less need to counteract small conditional power values. Additionally, one can also observe that for smaller effect sizes, for example, $\delta = 0.3$, the sample size curve of the point prior tends to the sample size curve of a classic group-sequential design (data not shown). For the two continuous priors, the shift towards smaller effect sizes results in a sample size increase to attain reasonable conditional power values (data not shown). Score values for further values of $\delta$ are reported in Appendix B.

It is important to mention that in addition to the presented examples of a point, a normal, and a uniform prior, there are many other options how the trial prior can be chosen. In general, there is no one-size-fits-all approach and the prior choice always depends on specific aspects of the particular clinical trial. Various literature discussing this issue and giving detailed recommendations exists (Dallow et al., 2018; Rufibach et al., 2016; Spiegelhalter et al., 2004), and we would like to refer to these publications for support with prior elicitation.

5 | WEIGHTING OF SCORE COMPONENTS

5.1 | Altered weighting schemes

In Section 4, we observed that two designs might seem similar in performance while the score components substantially deviate. The question is thus if all components should be weighted equally. Clearly, there exist other possibilities than an equal weighting ($S_0$) of the four components in the conditional performance score. As we have seen in Table 1, not all score components perform equally well and in a particular setting there might be a substantial reason for not weighting all four components equally. A higher weighting of both variation components $s_{v,n}$ and $s_{v,cp}$ compared to the location components $s_{e,n}$ and $s_{e,cp}$ seems not meaningful since it favors designs with a nearly constant stage-two sample size and at the same time small stage-two sample size. However, it can be reasonable to attribute higher weights to the location components $s_{e,n}$ and $s_{e,cp}$. Therefore, we investigate the additional three weighting scenarios $S_1$, $S_2$, $S_3$ as given in Table 2.

Moreover, it might also be useful to react to observed component values in the $S_0$ scenario. It can be seen from Table 1 that the $s_{e,n}$-component performance is almost perfect while the $s_{e,cp}$-component performance is rather poor. Therefore, weighting scenario $S_4$ is also included in Table 2 to give the $s_{e,cp}$-component a higher weight than the $s_{e,n}$-component.

These four new weighting schemes ($S_1$, $S_2$, $S_3$, $S_4$) are applied to the point, truncated normal, and uniform prior distribution from before ($\pi_{\text{point}}$, $\pi_{\text{normal}}$, $\pi_{\text{uniform}}$). The resulting performance with respect to the conditional
TABLE 2  Weights of the score components \((\omega_{e,n}, \omega_{v,n}, \omega_{e,cp}, \omega_{v,cp})\) resulting from different weighting schemes. The weights are independent of the chosen prior.

| Weighting scheme | \(\omega_{e,n}\) | \(\omega_{v,n}\) | \(\omega_{e,cp}\) | \(\omega_{v,cp}\) |
|------------------|-----------------|-----------------|-----------------|-----------------|
| \(S_0\)          | 0.25            | 0.25            | 0.25            | 0.25            |
| \(S_1\)          | 0.3             | 0.2             | 0.3             | 0.2             |
| \(S_2\)          | 0.35            | 0.15            | 0.35            | 0.15            |
| \(S_3\)          | 0.4             | 0.1             | 0.4             | 0.1             |
| \(S_4\)          | 0.3             | 0.1             | 0.5             | 0.1             |

TABLE 3  Evaluation of conditional performance score and its components \((s_{e,n}, s_{v,n}, s_{e,cp}, s_{v,cp})\) in dependence of different weighting schemes and of prior choice.

| Prior         | Weighting scheme | Optimal design | \(s_{e,n}\) | \(s_{v,n}\) | \(s_{e,cp}\) | \(s_{v,cp}\) | Score     |
|---------------|------------------|----------------|------------|------------|------------|------------|-----------|
| Point         | \(S_1\)          | \(D_{point,S_1}\) | 0.99962    | 0.93380    | 0.59866    | 0.52417    | 0.77090   |
| Point         | \(S_2\)          | \(D_{point,S_2}\) | 0.93778    | 1.00000    | 0.72066    | 0.33745    | 0.78098   |
| Point         | \(S_3\)          | \(D_{point,S_3}\) | 0.92026    | 0.94043    | 0.75668    | 0.34039    | 0.79869   |
| Point         | \(S_4\)          | \(D_{point,S_4}\) | 0.79262    | 0.93812    | 0.86326    | 0.31089    | 0.79458   |
| Normal        | \(S_1\)          | \(D_{normal,S_1}\) | 0.99968    | 0.83502    | 0.35139    | 0.75009    | 0.72258   |
| Normal        | \(S_2\)          | \(D_{normal,S_2}\) | 0.99928    | 0.83549    | 0.35317    | 0.74423    | 0.71064   |
| Normal        | \(S_3\)          | \(D_{normal,S_3}\) | 0.99973    | 0.83793    | 0.35347    | 0.73985    | 0.69918   |
| Normal        | \(S_4\)          | \(D_{normal,S_4}\) | 0.99973    | 0.83706    | 0.35386    | 0.74028    | 0.63463   |
| Uniform       | \(S_1\)          | \(D_{uniform,S_1}\) | 0.99952    | 0.71601    | 0.39339    | 0.68872    | 0.69889   |
| Uniform       | \(S_2\)          | \(D_{uniform,S_2}\) | 0.99982    | 0.71546    | 0.39464    | 0.68426    | 0.69809   |
| Uniform       | \(S_3\)          | \(D_{uniform,S_3}\) | 0.99998    | 0.71506    | 0.39524    | 0.68217    | 0.69784   |
| Uniform       | \(S_4\)          | \(D_{uniform,S_4}\) | 0.99968    | 0.71318    | 0.39595    | 0.68112    | 0.63740   |

performance score as well as its four components are listed in Table 3, and the corresponding plots can be found in Figure 4.

Regarding the point prior distribution \(\pi_{point}\), all four new weighting schemes improve the \(s_{e,cp}\)-component values compared to the \(S_0\)-weighting scheme. This, however, comes at the price of smaller \(s_{v,cp}\)-component values. No weighting scheme leads to a good performance in all four components. The sample size curves are more bended the lower the weight for the variance components becomes. Naturally, when giving more weight to the conditional power location component, higher sample sizes are attained (Figure 4). Similarly, the conditional power curve values increase with higher sample size values.

With respect to the continuous prior distributions (normal prior distribution \(\pi_{normal}\) and uniform prior distribution \(\pi_{uniform}\)), the score component values as well as the total score values are rather similar across the different weightings per considered prior distribution. The sample size curves for the new weighting scenarios \((S_1, S_2, S_3, S_4)\) differ only slightly from each other (Figure 4). Note, however, that certain weightings lead to wavy sample size curves, which cannot be recommended in application.

We have seen that the different performance score component weightings do not or only slightly increase the performance score value. The biggest influence of the weighting constellation could be found for the point prior to \(\pi_{point}\). Furthermore, the resulting sample size function shape is of utmost importance. Another strategy to possibly improve the overall performance could be given by a more flexible optimization setting, for example, with a flexible critical value function. Moreover, good performance with respect to a specific performance measure can be addressed by adding additional constraints to the optimization problem. This will be addressed in Section 6.

5.2  Inclusion of score components in the optimization problem

Besides developing rules for different weighting schemes, the four score components \(s_{e,n}, s_{v,n}, s_{e,cp}, s_{v,cp}\) can also be included in the optimization problem.
5.2.1 Change of the objective function

The first presented approach to achieve a sufficiently large performance in all score components is not to optimize over the total score $S$ but to maximize the minimum of the four score components. Thus, the problem to solve is

$$\maximize_{n_Z} \min \{ s_{d,\text{lin}}(D, \delta), s_{d,\text{lin}}(D, \delta), s_{e,\text{lin}}(D, \delta), s_{e,\text{lin}}(D, \delta) \}. \tag{9}$$

This modified objective criterion ensures that a large value is achieved in all subscores by increasing each component as much as possible without strongly decreasing the other components. Thus, the acceptable performance of all score components is anticipated.

We solve this problem for the point, the normal, and the uniform prior that were introduced in Section 4. In Figure 5, the three resulting designs are depicted. We observe that the designs do not differ strongly in dependence of the underlying prior. Compared with the designs solving the initial optimization problem, the shape of the design under the point prior has hardly changed (cf. Figure 1), while the two continuous priors yield more variable sample size functions (cf. Figure 2). This causes an increased value of the conditional power components compared to the initial optimization problem (cf. Table 4).
Table 4 shows the resulting score components of the three designs. Interestingly, for all three priors, all components are increased above a level of 0.5. For the point prior, the lowest values are observed for the two conditional power subscores but both lie above a boundary of 0.64. While for the original optimization problem, the lowest subscore for the two continuous priors was the expectation component of the conditional power (cf. Table 3), now both conditional power components are at a similar level of about 0.56 for the normal and 0.51 for the uniform prior.

5.2.2 Constraints on score components

In the previous section, we observed that maximizing the minimum of the four score components yields an acceptable performance in all these components. However, the overall score was no longer considered within the optimization problem. A compromise between overall score performance and performance in the four components may be to maximize the total score but to impose constraints on the components. This ensures that the best possible score value is achieved while a sufficient performance of all components is guaranteed.

As an example, we consider the case of equal weighting, that is \( \omega_{e,n} = \omega_{v,n} = \omega_{e,cp} = \omega_{v,cp} = 0.25 \). The optimization problem to solve is

\[
\text{maximize} \quad S_0(D, \delta) \\
\text{subject to} \quad s_{e,n} \geq lb_{e,n}, \quad s_{v,n} \geq lb_{v,n}, \quad s_{e,cp} \geq lb_{e,cp}, \quad s_{v,cp} \geq lb_{v,cp}.
\]

Again, this problem is solved for the point, the normal, and the uniform prior introduced in Section 4. For the point prior, we choose \( lb_{e,n} = lb_{v,n} = lb_{e,cp} = lb_{v,cp} = 0.6 \) and for the two continuous priors, the boundaries \( lb_{e,n} = lb_{v,n} = lb_{e,cp} = lb_{v,cp} = 0.5 \) are chosen.

In Figure 6, the three resulting designs are presented. The sample size function under the two continuous priors is not very volatile and is very similar for both priors. For the point prior, the resulting design is similar to the optimal design from Section 3 without constraints on the score components. For the normal and the uniform priors, the designs do hardly differ. Comparing the continuous priors with the optimization without constraints (Figure 2) and the “max-min” approach (Figure 5), the sample sizes are not as volatile as in the latter case and similar as volatile as in the first case.
The performance in the score components is reported in Table 5. We observe that the constraints on the subscores are indeed fulfilled in all three cases. Comparing the overall score value with the score maximization without additional constraints, we see a slight performance deterioration (cf. Table 1). If we compare this constraint-based approach with the idea of maximizing the minimum of the four components, an improvement in the overall score is observed (cf. Table 4). This confirms the thesis that this approach may be a valid compromise between overall and subscore performance.

6 EXTENDING THE OPTIMIZATION APPROACH

In this section, we describe how the score optimization approach can be extended. This will be done in two manners. First, we describe in Section 6.1 how a design can be tailored by requesting additional constraints. Second, we analyze in Section 6.2 the approach to not only optimize the stage-two sample size function \( n_2(z) \) but also the stage-two critical value function \( c_2(z) \). To illustrate the procedures in this section, we return to the initially optimized score \( S_0 \) with an equal component weighting and design \( D_1 \) whose sample size function maximizes \( S_0 \). It is depicted in Figure 1, and it achieves a score value of \( S_0(D_1, \delta = 0.4) = 0.76835 \).

6.1 Adding constraints

The proposed score is conditional in the sense that it only considers first-stage test statistics falling into the continuation region, that is, \( z_1 \in [c_f, c_c] \). As demonstrated above, this helps to choose an appropriate sample size recalculation rule, that is, an appealing function \( n_2(z_1) \) for \( z_1 \in [c_f, c_c] \). However, unconditional properties are highly relevant as well even though the adaptive design allows modifications during the interim analysis. Without losing the option to adapt the trial design, a two-stage design \( D \) can be prespecified during the planning stage by defining its elements \( n_1, c_f, c_c, n_2(\cdot), \) and \( c_2(\cdot) \). While the overall type I error rate is protected by appropriately chosen critical values for both stages, further unconditional elements may be considered as well.

As an exemplary unconditional constraint, we consider the unconditional power that is defined as the overall probability to reject the null hypothesis if the alternative is true. It is also a highly relevant characteristic of a multistage design. Even though the unconditional power is no element of the conditional performance score by Herrmann et al. (2020), it can be included in the setting of this paper as well. For this purpose, note that a power constraint can also be fulfilled by a suitable choice of the stage-two sample size \( n_2(\cdot) \). This can easily be achieved by incorporating a constraint into the optimization.
problem that the unconditional power should be at least $1 - \beta$. The optimization problem to solve would read as

$$\begin{align*}
\text{maximize} & \quad S_0(D, \delta = 0.4) \\
\text{subject to} & \quad \text{unconditional power at } \delta = 0.4 \geq 1 - \beta.
\end{align*}$$

(11)

This problem can be solved analogously to the previously defined optimization problems and guarantees a high score value and a sufficiently large unconditional power simultaneously. Note, however, that the optimal score value may decrease by imposing this additional constraint.

In our example, the optimal design $D_1$ achieves an unconditional power of 73.7% for the point effect $\delta = 0.4$ and thus does not hit the target value of $1 - \beta = 80\%$. Consequently, one may want to modify the design in order to guarantee the power condition by solving the optimization problem in (11). In Figure 7, the resulting design is plotted (dashed line).

$D_4$, the optimized design under additional global power constraint, hits the power target of $1 - \beta = 80\%$ for $\delta = 0.4$. The shape of $n_2(\cdot)$ is quite similar to design $D_1$, that is, it is increasing for values of $z_1$ that are close to the futility boundary $c_f$ and decreasing for larger values of $z_1$. To fulfill the power constraint, slightly larger sample sizes are needed within the entire continuation region. The additional constraint implies that the score value decreases to $S_0(D_4, \delta = 0.4) = 0.75688$.

The unconditional power is not the only possible constraint that can be included in the optimization procedure. Many other options are imaginable. For instance, one may also consider lower boundaries for particular subscores of $S$ to ensure that the resulting design shows an acceptable performance in all score components. Furthermore, one could also include lower or upper boundaries for the conditional power as it is already done for the maximum sample size that must not exceed the value of $n_{\text{max}}$. In the next section, we include the maximum type I error rate as constraint and use this to optimize the critical value function $c_2(\cdot)$ as well. Note that this was not necessary up to now because the choice of $c_2(\cdot)$ according to a predefined combination test ensured type I error rate protection.

6.2 Optimization of the critical value function

The conditional score $S$ is not only influenced by the sample size function $n_2(\cdot)$ but also by the stage-two critical value $c_2(\cdot)$. This holds due to the fact that the conditional power components of $S$ do also depend on $c_2(\cdot)$. Therefore, one may use $S$ not only to judge different sample size choices but also to compare different expressions of the critical value function $c_2(\cdot)$. This idea is explored in this section.

Up to now, $c_2(\cdot)$ was chosen in accordance with the inverse normal combination test throughout this paper. The first proposal of a combination test mentioned in the literature on adaptive designs did not use the nowadays frequently used inverse normal method but Fisher’s $p$-value combination method (Bauer & Köhne, 1994). Thus, the null hypothesis is rejected at the final analysis if the product of the stage-wise $p$-values falls below a predefined boundary $c_2$. In the notation

![Figure 7: Optimal designs with and without power constraint](image-url)
of this paper, the corresponding critical value function is given as

\[ c_2(z_1) = \Phi^{-1}\left( 1 - \frac{c_\alpha}{1 - \Phi(z_1)} \right), \quad \text{(12)} \]

where \( \Phi \) denotes the cumulative distribution function of the standard normal distribution. Instead of protecting the maximum type I error rate by means of the inverse normal method, one may also use Fisher’s product test. As outlined above, the conditional performance score by Herrmann et al. (2020) is not restricted to the inverse normal method but can be applied to any kind of combination test. Therefore, it can also be applied to compare different choices of \( c_2(\cdot) \). To allow a fair comparison, it is important to restrict to the same recalculation area \([c_f, c_e]\) when comparing different critical value functions. Therefore, in the following, only the stage-two critical value \( c_2(\cdot) \) is altered while the continuation region is not modified.

In concordance with the previous considerations, one may also be interested in computing the optimal stage-two sample size \( n_2(\cdot) \) with respect to \( S_0 \) when Fisher’s product test is used. In Figure 8, the resulting design \( D_5 \) is illustrated (dashed line). The optimal sample size function appears to be similar to the inverse normal case. It is also unimodal but slightly flatter than the inverse-normal-based one. This implies that the maximum sample size is slightly smaller, and the sample sizes close to the early stopping boundaries are slightly increased. The stage-two critical value function \( c_2(\cdot) \) is concave according to the definition in Equation (12). While the design \( D_1 \) under the inverse normal combination test achieves a score value of \( S_0(D_1, \delta = 0.4) = 0.76835 \), the design under Fisher’s product test achieves a value of \( S_0(D_5, \delta = 0.4) = 0.77291 \).

Besides using a predefined combination test, one may also apply the function \( c_2(\cdot) \) that is maximizing \( S_0 \). In this case, type I error rate control is not ensured and must be added as further constraint. The optimization problem to solve becomes

\[
\begin{align*}
\text{maximize} & \quad S_0(D, \delta = 0.4) \\
\text{subject to} & \quad \text{type I error rate} \leq \alpha.
\end{align*}
\quad \text{(13)}
\]

The design solving (13) is shown in Figure 8 (dot-dashed line). The resulting optimal \( c_2(\cdot) \) function is decreasing for values close to the futility boundary but increases for larger values of \( z_1 \). The large values of \( c_2(\cdot) \) close to the early futility boundary may indicate that a more aggressive futility stop would be favored by the optimization procedure since for those large critical stage-two values, rejection of the null hypothesis at the final analysis is very unlikely. The non-monotonic curve may not be acceptable in practice since larger values of \( z_1 \) that indicate more evidence against the null hypothesis should indicate smaller values of \( c_2(z_1) \) to facilitate the rejection of \( H_0 \). Consequently, the optimized critical value function cannot be recommended. It achieves a score value of \( S_0(D_7, \delta = 0.4) = 0.81375 \).
As a consequence, one may optimize over $c_2(\cdot)$ under the condition that it must be monotonically decreasing. This yields the optimization problem:

\[
\begin{align*}
\text{maximize} & \quad n_2(\cdot)c_2(\cdot) \\
\text{subject to} & \quad S_0(D, \delta = 0.4) \\
& \quad \text{type I error rate } \leq \alpha \\
& \quad c_2(\cdot) \text{ is monotonically decreasing.}
\end{align*}
\] (14)

The resulting design is shown in Figure 8 (dotted line). In comparison to an arbitrary shape of $c_2(\cdot)$, the stage-two critical value is now constant in the area where it has been increasing before. The steep decrease for values of $z_1$ close to $c_f$ remains identical and again implies a very low conditional power. The corresponding optimal sample size function is no longer constant but unimodal. The score value of this design equals $S_0(D_{\text{opt}}, \delta = 0.4) = 0.79645$ and is thus smaller than that of design $D_f$ with more degrees of freedom but slightly increased as compared to the inverse normal and the Fisher design whose critical value function is monotonic but not optimized.

Various facets of how to extend this approach are imaginable. For instance, in the case of group-sequential designs, one may optimize over optimum-delta critical value functions (Wang & Tsiatis, 1987). The usage of more or less steep conditional error functions in order to investigate the potential score maximum is as well imaginable. Furthermore, one may investigate the influence of the underlying effect size $\delta$ and the score weighting scheme. Some results for the latter points are provided in Appendix B.

7 | DISCUSSION

Within this article, we have introduced and evaluated optimal adaptive two-stage designs with respect to the conditional performance score proposed by Herrmann et al. (2020). This optimization was conducted to address two different purposes: (1) The optimization approach determines the optimal sample size curve when the conditional performance score is used as the scoring criterion. (2) For a specific setting, the optimization approach with respect to the conditional performance score determines the best possible conditional score value. This can serve as a realistic upper score boundary when comparing different sample size recalculation rules in that specific setting with respect to the score under a specific prior. We considered the conditional performance score optimization for different prior distributions, subscore component weightings, and under further constraints.

Purpose 1 could not be achieved unequivocally. We have only observed slight improvements in the conditional performance score when comparing an optimal adaptive design with a group-sequential design. As expected, continuous prior distributions cause flatter optimal sample size curves on average compared to a point prior distribution. Nevertheless, especially the uniform prior distribution may lead to wavy sample size curves for certain parameter constellations, which is difficult to communicate in application. Of note, the incorporation of the continuous prior distribution is a beneficial extension of the point-prior case since it adequately addresses the insecurity about the effect size. Furthermore, different score component weighting schemes do not or only slightly increase the conditional performance score value and lead to very similar sample size curves for the considered continuous prior distributions. Clearly, there exist more weighting possibilities for the subscores than the ones presented within this paper.

Considering purpose 2, we have seen that the optimization approach can be used to calculate actual upper benchmarks for the conditional performance score value under given design assumptions. Within this paper, we have seen no optimal design with a conditional performance score value of 1. Therefore, comparing the score value of an initially planned design with the best possible value may help to decide whether this design can be deemed acceptable for application. It is important to mention that an exclusive look at the difference between two score values may not yield all delivered information. Obviously, higher score values indicate better performance. However, it is difficult to interpret a score difference of, for instance, 0.01 between two designs because one has also to consider the maximally achievable score value in this setting. Thus, a relative comparison of the achieved score of a certain design with the maximally achievable score value may often be more reasonable than an exclusive analysis of the absolute score value difference. Specific unconditional performance measures can be addressed by additional constraints, and applying Fisher’s product test can improve the overall performance score value. The optimization of the critical value function, however, cannot be recommended in the considered setting due to a partially increasing shape for increasing interim test statistics.
It should be noted that due to the complex character of the investigated score, a stable optimization is not straightforward. The general methodology to optimize the score $S$ was implemented in concordance with the concept of the R-package `adoptr` (Kunzmann et al., 2021). Following the optimization approach of `adoptr`, the continuation region $[c_f, c_c]$ was split by $k$ nodes. During our research, we experienced that a too large number of nodes leads to decreased score values of the optimized designs when it is integrated over the $[c_f, c_c]$ region. Therefore, we decided to use six nodes within this paper. We think that a number between 5 and 7 leads to the most accurate results. A further aspect to consider is the starting point for the optimizer. As we experienced that the resulting optimal designs depend on the chosen starting design, the examples in this paper were prepared using different starting values for a global optimizer and continuing with the best performing design as starting value for a local optimizer. In addition, it cannot be ensured that each expression of the score $S$ is convex in the optimization parameters. These arguments imply that there exists no guarantee that the optimal design is always found. In particular, it can be observed (see, e.g., Section 4) that rather different designs show quite similar score values. This makes it difficult to recommend a specific design for practical application without further consideration. We recommend to optimize carefully, thus using different starting values, and to regard the resulting optimal design from different aspects (resulting function shape, magnitude of performance improvement compared to other sample size recalculation approaches, subscore performance etc.). These points become particularly pressing with increasing flexibility, that is increasing optimization parameters. The option to optimize over $c_2(\cdot)$ as well should thus be used with caution.

The interpretation of comparing score values resulting from different weighting schemes is of high relevance. The proposed score $S$ depends on the underlying prior distribution $\pi$ and the weight vector $\omega = (\omega_1, \omega_2, \omega_3, \omega_4)$. We think that, if these parameters vary, resulting score values cannot be compared directly. Regarding the frequently used scoring criterion “expected sample size”, it is also not meaningful to compare two adaptive designs by computing the expected sample size under effect size $\delta$ for the first and under effect size $\delta^* \neq \delta$ for the second design. The same argument holds for the score $S$. When two designs are to be compared and different weighting schemes or priors are used, all resulting scores have to be computed for both designs. In addition, a look at further design characteristics and the subscores is highly recommended to avoid designs that may not be applicable in practice (cf. Sections 5 and 6.2).

Furthermore, optimizing the score $S$ gives interesting insights into the structure of the score by detecting which components of the score may be weighted too strongly. In this paper, we observed that score-optimized designs often show a very high value of the component evaluating the expected sample size but quite small values of the component evaluating the conditional power. Therefore, deviating from the initially proposed equal weighting of all score components may be appropriate but should be done carefully and case driven. It is not possible to give clear guidance on how to choose the four weights $\omega_1, \omega_2, \omega_3, \omega_4$ in practice due to the many different aspects a clinical trial may have. However, we recommend to evaluate all subscores of a design and not to rely exclusively on the overall score value. If the subscore performance is not satisfactory, there are at least three options. One can modify the weighting scheme, include constraints on the subscores, or use another objective function. All these options have been presented in Section 5.

It is important to state that besides the prior and the weights, further elements of the score $S$ can be modified. For instance, one may use squared terms instead of absolute values in the score definition (cf. Appendix A) or use another combination of the score components than the linear combination that was initially proposed. An example of this idea is the minimum of the four subscores that have been presented in Section 5.2.1. In fact, each objective criterion can be optimized and there is no one-size-fits-all criterion that is perfectly suited to each clinical trial. Besides the frequently used expected sample size (under different effect sizes or priors), higher moments of the sample size function (Pilz et al., 2021), a linear combination of power and expected sample size (Kunzmann & Kieser, 2020), or the maximum sample size (Schlömer & Brannath, 2013; Simon, 1989) have been proposed. All these criteria are at a certain point arbitrary, and it depends on the concrete clinical trial which criterion is deemed the most important one in this particular setting. Furthermore, all optimization problems and the resulting designs can be tailored by requiring additional constraints that may improve the design behavior in different aspects. However, users need to be aware of the fact that certain constraint combinations may lead to counterintuitive solutions.

Regarding communication towards clinicians of our method, we recommend pointing out the strength of combining different performance measures to determine optimality. This reflects the fact that evaluating a design is by far more comprising than adducing one measure. To allow a convenient evaluation and facilitate the presentation of the large number of options underlying the optimization approach, the authors provide an R shiny app that can be accessed under https://web.imbi.uni-heidelberg.de/recalculation-score/. It is important to mention that the above-described limitations of the presented optimization procedure also hold for this shiny app and that it should not be used as the unique basis for planning a clinical trial.
ACKNOWLEDGMENTS
This project has been funded by the Deutsche Forschungsgemeinschaft (DFG) under the grants KI 708/4-1, RA 2347/4-1, KI 708/4-2, and RA 2347/4-2.

The authors would like to thank the editor and three anonymous reviewers for their excellent comments that enormously helped to improve this manuscript.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST
The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT
Data was neither analyzed nor generated when preparing this paper. All R code that was used to compute the examples is available as supplemental material.

OPEN RESEARCH BADGES
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID
Carolin Herrmann https://orcid.org/0000-0003-2384-7303
Maximilian Pilz https://orcid.org/0000-0002-9685-1613

REFERENCES
Bartroff, J., & Lai, T. L. (2008). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. Statistics in Medicine, 27, 1593–1611.
Bauer, P., Bretz, F., Dragalin, V., König, F., & Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls. Statistics in Medicine, 35, 325–347.
Bauer, P., & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. Biometrics, 50, 1029–1041.
Bauer, P., & König, F. (2006). The reassessment of trial perspectives from interim data—A critical view. Statistics in Medicine, 25, 23–36.
Brannath, W., & Bauer, P. (2004). Optimal conditional error functions for the control of conditional power. Biometrics, 60, 715–723.
Chow, J. T., Tsiatis, A. (2008). Sample size calculations for randomized clinical trials published in anesthesiology journals: A comparison of 2010 versus 2016. Canadian Journal of Anesthesia/Journal canadien d'anesthésie, 65, 611–618.
Dallow, N., Best, N., & Montague, T. (2018). Better decision making in drug development through adoption of formal prior elicitation. Pharmaceutical Statistics, 17, 301–316.
Dallow, N., & Fina, P. (2011). The perils with the misuse of predictive power. Pharmaceutical Statistics, 10, 311–317.
Fang, F., Lin, Y., Shish, W. J., Lu, S.-E., & Zhu, G. (2018). Evaluation of performance of adaptive designs based on treatment effect intervals. International Journal of Statistics and Probability, 7, 81–93.
Herrmann, C., Pilz, M., Kieser, M., & Rauch, G. (2020). A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation. Statistics in Medicine, 39, 2067–2100.
Hsiao, S. T., Liu, L., & Mehta, C. R. (2019). Optimal promising zone designs. Biometrical Journal, 61, 1175–1186.
Jennison, C., & Turnbull, B. (2006). Adaptive and nonadaptive group sequential tests. Biometrika, 93, 1–21.
Jennison, C., & Turnbull, B. (2015). Adaptive sample size modification in clinical trials: Start small then ask for more? Statistics in Medicine, 34, 3793–3810.
Kunzmann, K., & Kieser, M. (2020). Optimal adaptive single-arm phase II trials under quantified uncertainty. Journal of Biopharmaceutical Statistics, 30, 89–103.
Kunzmann, K., Pilz, M., Herrmann, C., Rauch, G., & Kieser, M. (2021). The adoptr package: Adaptive optimal designs for clinical trials in R. Journal of Statistical Software, 98(9), 1–21.
Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. Biometrics, 55, 1286–1290.
Levin, G. P., Emerson, S. C., & Emerson, S. S. (2013). Adaptive clinical trial designs with pre-specified rules for modifying the sample size: Understanding efficient types of adaptation. Statistics in Medicine, 32, 1259–1275.
Liu, G. F., Zhu, G. R., & Cui, L. (2008). Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. Statistics in Medicine, 27, 584–596.
Lokhnygina, Y., & Tsiatis, A. A. (2008). Optimal two-stage group-sequential designs. Journal of Statistical Planning and Inference, 138, 489–499.
SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher’s website.

How to cite this article: Herrmann, C., Kieser, M., Rauch, G., & Pilz, M. (2022). Optimization of adaptive designs with respect to a performance score. Biometrical Journal, 64, 989–1006.
https://doi.org/10.1002/bimj.202100166

APPENDIX A: DEFINITION OF THE CONDITIONAL PERFORMANCE SCORE

The performance score by Herrmann et al. (2020) for an effect size \( \delta \) is given by

\[
S(D, \delta) := \omega_{e, n}s_{e, n}(D, \delta) + \omega_{v, n}s_{v, n}(D, \delta) + \omega_{e, cp}s_{e, cp}(D, \delta) + \omega_{v, cp}s_{v, cp}(D, \delta),
\]  
(A1)

with weights \( \omega_{e, n} + \omega_{v, n} + \omega_{e, cp} + \omega_{v, cp} = 1 \), where \( s_{e, n}(D, \delta), s_{e, cp}(D, \delta) \) refer to the location components and \( s_{v, n}(D, \delta), s_{v, cp}(D, \delta) \) refer to the variation components evaluating conditional power and sample size. More precisely, the four components are described as follows.

The sample size location component is given by

\[
s_{e, n}(D, \delta) = 1 - \mathbb{E}[N(Z_1) | Z_1 \in [c_f, c_e]] - n_{target, \delta} \bigg/ n_{max} - n_1,
\]  
(A2)

where \( \mathbb{E}[N(Z_1) | Z_1 \in [c_f, c_e]] \) describes the expected sample size conditional on entering the continuation region \([c_f, c_e]\) and the target sample size \( n_{target, \delta} \) is defined as the respective fixed design’s sample size at effect size \( \delta \) unless the underlying effect size \( \delta \) equals zero or the fixed design’s sample size exceeds the maximally feasible sample size \( n_{max} \), then it is given by \( n_1 \).

The sample size variation component is given by

\[
s_{v, n}(D, \delta) = 1 - \sqrt{\text{Var}(N(Z_1) | Z_1 \in [c_f, c_e]) / \text{Var}_{max}(N)},
\]  
(A3)
where \( \text{Var}(N(Z_1) \mid Z_1 \in [c_f, c_e]) \) describes the variance of the sample size conditional on entering the continuation region \([c_f, c_e]\), \( \text{Var}_{\text{max}}(N) \) is defined as \( ((n_{\text{max}} - n_1)/2)^2 \) and for a classical group sequential design it is defined as 0.

The conditional power location component is given by

\[
s_{e,\text{cp}}(D, \delta) = 1 - \left| \frac{\mathbb{E}[CP(Z_1) \mid Z_1 \in [c_f, c_e]] - CP_{\text{target}, \delta}}{1 - \alpha} \right|
\]

where \( \mathbb{E}[CP(Z_1) \mid Z_1 \in [c_f, c_e]] \) describes the expected conditional power conditional on entering the continuation region \([c_f, c_e]\) and the target conditional power \( CP_{\text{target}, \delta} \) is defined as \( 1 - \beta \) unless the underlying effect size \( \delta \) equals zero or the fixed design’s sample size exceeds the maximally feasible sample size, then it is given by \( \alpha \).

The conditional power variation component is given by

\[
s_{v,\text{cp}}(D, \delta) = 1 - \sqrt{\frac{\text{Var}(CP(Z_1) \mid Z_1 \in [c_f, c_e])}{\text{Var}_{\text{max}}(CP)}}
\]

where \( \text{Var}(CP(Z_1) \mid Z_1 \in [c_f, c_e]) \) describes the variance of the conditional power conditional on entering the continuation region \([c_f, c_e]\) and \( \text{Var}_{\text{max}}(CP) \) is defined as \((1 - 0)/2)^2 = 0.25.

Note that in the case of a continuous prior distribution \( \pi \), the score \( S(D, \pi) \) is defined by:

\[
S(D, \pi) := \int (\omega_{e,n}s_{e,n}(D, \delta) + \omega_{v,n}s_{v,n}(D, \delta) + \omega_{e,\text{cp}}s_{e,\text{cp}}(D, \delta) + \omega_{v,\text{cp}}s_{v,\text{cp}}(D, \delta))\pi(\delta) \, d\delta
\]

\[
= \omega_{e,n} \int s_{e,n}(D, \delta)\pi(\delta) \, d\delta + \omega_{v,n} \int s_{v,n}(D, \delta)\pi(\delta) \, d\delta + \omega_{e,\text{cp}} \int s_{e,\text{cp}}(D, \delta)\pi(\delta) \, d\delta + \omega_{v,\text{cp}} \int s_{v,\text{cp}}(D, \delta)\pi(\delta) \, d\delta.
\]

**APPENDIX B: COMPREHENSIVE SIMULATION**

In this section, a comprehensive simulation study is presented in order to depict the influence of the effect size \( \delta \), the stage-two critical value function, and the weightingscheme on the score value of the optimal design. For this purpose, we analyze values of \( \delta \in \{0.2, 0.4, 0.6, 0.8\} \). We present the critical value function \( c_2(\cdot) \) according to Fisher’s product test and the inverse normal combination test. Furthermore, a convex critical value function is investigated and defined as

\[
c_2^{\text{convex}}(z_1) = \exp(-z_1) \cdot 2.494 + 1.006
\]

that controls the type I error rate at level \( \alpha = 0.025 \) and shows a comparable range to the other two critical value functions. Additionally, the results of an optimized \( c_2(\cdot) \) are also reported. As weighting approaches, the case of equal weighting, the weighting scheme \( S_3 \) from Section 5.1 (\( \omega_{e,n} = 0.4, \omega_{v,n} = 0.1, \omega_{e,\text{cp}} = 0.4, \omega_{v,\text{cp}} = 0.1 \)), and the max-min case from Section 5.2.1 are investigated.

We report the value of the objective criterion (score or minimum of the score components, respectively) as well as the expected sample size under the respective effect size and the minimum and maximum conditional power of the designs. All values are reported in Table B1. We observe that the optimal designs strongly vary with the underlying combination function \( c_2(\cdot) \). This is expressed by the different values of the expected sample size and the conditional power. Regarding the weighting scheme, the smallest possible score values appear for the max-min objective, followed by the equal weighting of all score components, and the weighting schemes that favor the expectation components \( (S_3) \). These observations hold for all investigated values of \( \delta \). The larger the value of \( \delta \), the larger are the observed score values. The benefit of optimizing the critical value function \( c_2(\cdot) \) decreases with increasing \( \delta \). All these findings again point out the high dependence of the choice of underlying “hyperparameters” and thus the importance of a careful choice of all these design elements.
| $\delta$ | $c_2$-function | Weighting scheme | Objective value | $ESS_2$ | $CP_2(c_2)$ | $CP_2(c_1)$ |
|---|---|---|---|---|---|---|
| 0.2 | Fisher | Equal | 0.700 | 232.7 | 0.39 | 0.92 |
| 0.2 | Inverse normal | Equal | 0.692 | 234.1 | 0.21 | 0.91 |
| 0.2 | Convex | Equal | 0.701 | 237.7 | 0.12 | 0.84 |
| 0.2 | Optimal | Equal | 0.819 | 297.6 | 0.05 | 0.01 |
| 0.2 | Fisher | S3 | 0.840 | 297.4 | 0.57 | 0.94 |
| 0.2 | Inverse normal | S3 | 0.818 | 297.0 | 0.32 | 0.81 |
| 0.2 | Convex | S3 | 0.829 | 296.9 | 0.10 | 0.82 |
| 0.2 | Optimal | S3 | 0.872 | 297.6 | 0.00 | 0.29 |
| 0.2 | Fisher | Max-min | 0.565 | 167.5 | 0.22 | 0.35 |
| 0.2 | Inverse normal | Max-min | 0.549 | 169.0 | 0.09 | 0.29 |
| 0.2 | Convex | Max-min | 0.550 | 158.7 | 0.03 | 0.24 |
| 0.2 | Optimal | Max-min | 0.591 | 175.3 | 0.00 | 0.42 |
| 0.4 | Fisher | Equal | 0.773 | 76.8 | 0.36 | 0.28 |
| 0.4 | Inverse normal | Equal | 0.768 | 76.8 | 0.12 | 0.34 |
| 0.4 | Convex | Equal | 0.779 | 76.8 | 0.06 | 0.52 |
| 0.4 | Optimal | Equal | 0.814 | 76.8 | 0.00 | 0.25 |
| 0.4 | Fisher | S3 | 0.794 | 100.4 | 0.39 | 0.93 |
| 0.4 | Inverse normal | S3 | 0.799 | 96.4 | 0.17 | 0.92 |
| 0.4 | Convex | S3 | 0.804 | 101.5 | 0.18 | 0.91 |
| 0.4 | Optimal | S3 | 0.805 | 96.7 | 0.00 | 0.85 |
| 0.4 | Fisher | Max-min | 0.646 | 107.3 | 0.68 | 0.60 |
| 0.4 | Inverse normal | Max-min | 0.643 | 104.3 | 0.51 | 0.46 |
| 0.4 | Convex | Max-min | 0.639 | 103.3 | 0.95 | 0.55 |
| 0.4 | Optimal | Max-min | 0.676 | 114.0 | 0.15 | 0.54 |
| 0.6 | Fisher | Equal | 0.782 | 51.7 | 0.99 | 1.00 |
| 0.6 | Inverse normal | Equal | 0.794 | 50.9 | 0.00 | 0.20 |
| 0.6 | Convex | Equal | 0.800 | 58.6 | 0.00 | 0.68 |
| 0.6 | Optimal | Equal | 0.810 | 71.6 | 0.96 | 1.00 |
| 0.6 | Fisher | S3 | 0.820 | 70.2 | 0.62 | 1.00 |
| 0.6 | Inverse normal | S3 | 0.825 | 72.0 | 0.70 | 1.00 |
| 0.6 | Convex | S3 | 0.827 | 70.3 | 0.14 | 1.00 |
| 0.6 | Optimal | S3 | 0.831 | 81.2 | 0.99 | 1.00 |
| 0.6 | Fisher | Max-min | 0.630 | 78.8 | 0.99 | 1.00 |
| 0.6 | Inverse normal | Max-min | 0.619 | 52.8 | 0.93 | 1.00 |
| 0.6 | Convex | Max-min | 0.746 | 65.0 | 1.00 | 0.82 |
| 0.6 | Optimal | Max-min | 0.748 | 65.0 | 1.00 | 0.82 |
| 0.8 | Fisher | Equal | 0.824 | 53.1 | 1.00 | 1.00 |
| 0.8 | Inverse normal | Equal | 0.830 | 52.8 | 0.93 | 1.00 |
| 0.8 | Convex | Equal | 0.839 | 53.1 | 0.94 | 1.00 |
| 0.8 | Optimal | Equal | 0.850 | 53.8 | 0.96 | 1.00 |
| 0.8 | Fisher | S3 | 0.854 | 53.0 | 0.57 | 1.00 |
| 0.8 | Inverse normal | S3 | 0.860 | 52.8 | 0.44 | 1.00 |
| 0.8 | Convex | S3 | 0.858 | 53.1 | 0.74 | 1.00 |
| 0.8 | Optimal | S3 | 0.862 | 52.3 | 0.07 | 1.00 |
| 0.8 | Fisher | Max-min | 0.783 | 51.8 | 1.00 | 0.91 |
| 0.8 | Inverse normal | Max-min | 0.800 | 51.8 | 1.00 | 0.90 |
| 0.8 | Convex | Max-min | 0.815 | 52.0 | 1.00 | 0.97 |
| 0.8 | Optimal | Max-min | 0.816 | 52.0 | 1.00 | 0.97 |