






Optimal planning of adaptive two-stage designs

Maximilian Pilz¹  | Kevin Kunzmann²  | Carolin Herrmann³  |
Geraldine Rauch³  | Meinhard Kieser¹ 

¹Institute of Medical Biometry and Informatics, University Medical Center Ruprecht-Karls University Heidelberg, Heidelberg, Germany

²MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Cambridge, UK

³Charité - Universitätsmedizin Berlin, Corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Berlin, Germany

Correspondence

Maximilian Pilz, Institute of Medical Biometry and Informatics, University Medical Center Ruprecht-Karls University Heidelberg, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany.
Email: pilz@imbi.uni-heidelberg.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Numbers: KI 708/4-1, RA 2347/4-1; RUPRECHT KARLS UNIVERSITAET HEIDELBERG - Projekt DEAL

Adaptive designs are playing an increasingly important role in the planning of clinical trials. While there exists various research on the optimal determination of a two-stage design, non-optimal versions still are frequently applied in clinical research. In this article, we strive to motivate the application of optimal adaptive designs and give guidance on how to determine them. It is demonstrated that optimizing a trial design with respect to particular objective criteria can have a substantial benefit over the application of conventional adaptive sample size recalculation rules. Furthermore, we show that in many practical situations, optimal group-sequential designs show an almost negligible performance loss compared to optimal adaptive designs. Finally, we illustrate how optimal designs can be tailored to specific operational requirements by customizing the underlying optimization problem.

KEYWORDS

adaptive design, clinical trial, optimal design, sample size calculation

1 | INTRODUCTION

The sensitive environment of medical research renders careful and responsible statistical planning of clinical trials necessary. An important task is the determination of the required sample size. Choosing the sample size too large may cause unnecessarily high costs and long time-to-market for potentially beneficial drugs. Vice versa, a too small sample size may imply that potentially underlying effects are not detected due to low power. This may lead to the erroneous and premature termination of development for actually beneficial drugs. Adaptive trial designs with their possibility to modify the sample size of an ongoing trial are an attractive option to decrease the average sample size while guaranteeing sufficient power. Within an adaptive design, one or several interim analyses are performed. At each interim analysis, it is decided whether the trial continues and how many further patients have to be recruited. This procedure implies that the sample

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

size within an adaptive design is a random variable depending on the interim results. Adaptive trial designs can thus be seen as tailoring the parameters of a design such that the characteristics of the random sample size are deemed agreeable. Recent overviews of the development of adaptive designs are given in Bauer et al¹ and, more detailed, in the book by Wassmer and Brannath.² In this article, we exclusively focus on the case of an adaptive two-stage design with one interim analysis.

A known issue in the analysis of adaptive two-stage designs is the fact that the stage-two data is not generally independent of the stage-one data. The methodology on how to test correctly within an adaptive design without type I error rate inflation is well-developed. Two main methods for ensuring strict type I error rate control in adaptive testing are put forward in the literature. The combination test approach³⁻⁵ exploits the fact that the stage-wise test statistics are identically distributed under the null hypothesis. It then reduces the stage-wise test statistic to a single overall test statistic using a so-called combination function. In contrast, the conditional error principle⁶ achieves type I error rate control by modifying the critical value for the final analysis. Both approaches have been shown to be equivalent for pre-planned interim analyses.⁷

Besides the various methods developed to assure type I error rate control, different proposals for the choice of the stage-wise sample sizes have been made. Initially, group-sequential designs⁸⁻¹⁰ were proposed to allow early termination of a trial for overwhelming good or bad interim outcomes. Since the second-stage sample size is constant for group-sequential designs, the power conditional on the observed outcome (conditional power) might drop below an acceptable threshold for some interim results. This drawback may make more flexible sample size rules that incorporate conditional power considerations desirable. Note that conditional power may be calculated in various ways, for example, inserting the treatment effect assumed in the planning phase, the interim estimate of the treatment effect, or integrated over the posterior distribution of the treatment effect. Throughout this article, we use the first approach and define conditional power as the probability to reject the null hypothesis under the initially assumed treatment effect conditioned on the interim result. The following considerations, however, hold for other definitions of conditional power as well. Requiring a minimal conditional power for each interim outcome implies that the sample size varies with the observed test statistic at interim.⁶ Bauer and König¹¹ analyzed the conditional power as random quantity varying with the stage-one outcome. The fact that the interim estimate of conditional power has a heavy-tailed distribution has severe consequences for early stopping rules. For instance, underestimating the true conditional power and thus the prospects of a successful second stage may lead to an incorrect early termination of the trial. Vice versa, overestimating the conditional power may imply that a second stage is conducted even though the chances of finally rejecting the null hypothesis are low. Further approaches based on conditional power such as the “promising zone design”¹² suffer from the same limitations and have in addition been criticized to be inferior to suitably chosen group-sequential designs.¹³

Besides these rules, suggestions have been put forward in the literature on the optimal determination of adaptive and group-sequential trial designs (cf. Section 2.2). Despite their theoretical appeal, optimal clinical trial designs are up to now rarely employed in practice. In this article, we discuss important aspects of optimal two-stage designs. In Section 2, we review the general optimization approach to trial design and existing approaches. We illustrate in Section 4 that the application of an optimal adaptive design may gain substantial benefit over the use of heuristic rules based on conditional power. Section 5 considers the issue of optimal group-sequential designs being an appropriate approximation to optimal adaptive ones even if a constraint on conditional power is included. In Section 6, we demonstrate how to incorporate additional constraints to tailor optimal designs to a specific situation. We conclude with a discussion in Section 7.

2 | OPTIMAL PLANNING OF ADAPTIVE DESIGNS

2.1 | Setting

Throughout this article, we consider the situation of a treatment group T to be compared against a control group C . The outcomes of interest, X_i^T and X_i^C , are assumed to be normally distributed with mean values μ_T and μ_C and common known variance σ^2 . The superiority null hypothesis on the mean difference θ is

$$\mathcal{H}_0 : \theta := \mu_T - \mu_C \leq 0.$$

In the following, we denote the standardized mean difference by $\delta := (\mu_T - \mu_C)/\sigma$ and introduce an effect size $\delta_1 > 0$ for which power constraints are required to be fulfilled.

To construct a two-stage design, assume that n_1 patients per group are to be included in the first and n_2 patients per group in the second stage. Then, the test statistics for the two stages are given by

$$Z_i = \sqrt{n_i} \frac{\bar{X}_{n_i}^T - \bar{X}_{n_i}^C}{\sqrt{2\sigma^2}}, \quad i \in \{1, 2\},$$

where \bar{X}_n denotes the mean value of n observations. At the interim analysis, it has to be decided whether the trial should be stopped early. Therefore, early stopping boundaries c_f and c_e are introduced where the trial is stopped for futility if $Z_1 < c_f$ and for efficacy if $Z_1 > c_e$. Otherwise, if $c_f \leq Z_1 \leq c_e$, the trial enters the second stage and $n_2(Z_1)$ further patients per group are recruited. At the final analysis, the null hypothesis is rejected if $Z_2 \geq c_2(Z_1)$. An adaptive two-stage design can thus be seen as a five-tuple $\mathcal{D} := (n_1, c_f, c_e, n_2(\cdot), c_2(\cdot))$ where the stage-two parameters $n_2(\cdot)$ and $c_2(\cdot)$ are functions of the first-stage outcome $Z_1 = z_1$. This definition corresponds with the concept of “planned flexible designs”¹⁴ where all decisions regarding the trial design are already specified during the planning stage. Note that the presented methodology can be applied to all (asymptotically) Gaussian distributed test statistics. Examples for such tests are the t -test, the log-rank test, and the two-sample binomial test.

A commonly applied strategy on how to plan an adaptive two-stage design is the following. In the planning phase, the first-stage elements n_1 , c_f , and c_e are fixed. To ensure type I error rate protection, a combination function $C(Z_1, Z_2)$ (or, equivalently, a conditional error function) is specified a priori. At the final analysis, the test statistics Z_1 and Z_2 are combined via C . These approaches control the maximum type I error rate. However, the choice of design parameters is not guided by a trial-specific performance criterion.

To define an efficient design, a strategy on how to identify all parameters of a two-stage design should be employed. Here, we define efficiency in terms of a performance criterion $f(\mathcal{D})$ mapping any particular design to a numeric performance value. In the following, we assume smaller values of f to indicate good performance. The optimal design $\mathcal{D}^* := (n_1^*, c_f^*, c_e^*, n_2^*(\cdot), c_2^*(\cdot))$ is then given by the design minimizing f potentially under a set of trial-specific constraints. Hence, it is defined by the solution of the constrained optimization problem

$$\underset{\mathcal{D}}{\text{minimize}} \quad f(\mathcal{D}) \quad (1)$$

$$\text{subject to} \quad g_1(\mathcal{D}) \leq a_1 \quad (2)$$

...

$$g_k(\mathcal{D}) \leq a_k, \quad (3)$$

where g_i , $i \in \{1, \dots, k\}$, are performance criteria. Note that although the optimal design's sample size function is response-adaptive, all parameters are pre-specified during the planning stage. Therefore, the perspective of deriving an optimal design before any data is collected is inherently unconditional since all possible interim outcomes have to be considered simultaneously. Consequently, the objective criterion should also be unconditional, that is, integrating over all potential interim outcomes weighted with their respective likelihood.

Constraints, however, may be both unconditional as well as conditional on the observed outcome $Z_1 = z_1$. Important unconditional constraints are maximum type I error rate, minimal power, or maximum sample size. Furthermore, desirable conditional properties such as a threshold on minimal conditional power can be incorporated for all values $z_1 \in [c_f, c_e]$. This may help in making a design more appealing for a sponsor since it only enters a second stage with a sufficient chance of detecting a truly existing effect. However, any further constraint reduces the solution space of the optimization problem (1)-(3). Hence, the unconditional performance is deteriorated by any additional constraints that are binding in the optimal solution. This interplay will be illustrated in Section 6.3. In particular, it may even occur that the solution space of (1)-(3) is empty. For instance, a strict constraint on the maximum sample size may imply that a certain power boundary cannot be reached. It is in the statistician's responsibility to choose criteria that are not so restrictive or even contradictory that a design fulfilling all these criteria may show disadvantageous features or may not exist at all.

2.2 | Previous work

Since $n_2(\cdot)$ and $c_2(\cdot)$ are functions, the task of computing optimal $n_2^*(\cdot)$ and $c_2^*(\cdot)$ is a variational problem. There are various approaches to tackle this problem and different ways have been pursued to solve specific variants of the problem (1)-(3).

Different excellent publications solved the problem for one of the two functions $n_2(\cdot)$ and $c_2(\cdot)$ exclusively. Optimal group-sequential designs, that is, designs with constant stage-two sample size function $n_2(\cdot) \equiv n_2$, were proposed by Barber and Jennison.¹⁵ They minimized a mixed criterion consisting of expected sample size under different effect sizes and derived optimal group-sequential tests, that is, optimal functions $c_2^*(\cdot)$. Note that they provide their results not only for the two-stage case but in more generality for $K \geq 2$ stages. Brannath and Bauer¹⁶ derived the formal computation of an optimal conditional error function for a sample size rule based on conditional power.⁶ They obtained $c_2^*(\cdot)$ by minimizing the expected sample size under the alternative $\delta = \delta_1$ while the sample size function was defined by

$$n_2(z_1) = 2 \cdot \frac{(c_2^*(z_1) + \Phi^{-1}(1 - \beta_c))^2}{\delta_1^2}, \quad (4)$$

where Φ denotes the cumulative distribution function of the standard normal distribution. As outlined in the introduction, this rule guarantees that after observing z_1 at the interim analysis, a conditional power of $1 - \beta_c$ is reached under the initially applied alternative effect size $\delta = \delta_1$. Jennison and Turnbull¹³ approached the problem from the sample size perspective and suggested an optimal $n_2^*(\cdot)$ fixing the conditional error function by an inverse normal combination test. Hence, $c_2(\cdot)$ was chosen as

$$c_2(z_1) = \frac{c - \omega_1 z_1}{\omega_2}, \quad (5)$$

where c is defined to protect the type I error rate and $\omega_1, \omega_2 \in (0, 1)$ are chosen such that $\omega_1^2 + \omega_2^2 = 1$. Then, the stage-two sample size $n_2^*(\cdot)$ was determined to minimize the expected sample size under the alternative under a constraint on the design's overall power. Hsiao et al¹⁷ proposed a design that shows similar performance as the design by Jennison and Turnbull by incorporating a constraint that the stage-two sample size can only be increased if a certain conditional power condition is satisfied.

Various extensions exist which are providing optimal $n_2^*(\cdot)$ and $c_2^*(\cdot)$ simultaneously. "Optimal sequentially-planned sequential tests" were already proposed in an abstract form by Schmitz.¹⁸ This proposal was applied by Jennison and Turnbull^{19,20} who regarded the problem to derive optimal designs from the perspective of Bayesian decision theory and solved it by the backward induction algorithm. They derived optimal adaptive designs for $K \geq 2$ stages which minimize the integral of expected sample size over a normal density for the treatment effect δ .¹⁹ Furthermore, they figured out that the efficiency gain produced by optimal adaptive designs in comparison with optimal group-sequential designs is quite small.¹⁹ Another result by Jennison and Turnbull²⁰ is the sound performance of well-chosen ρ -family error-spending tests which achieve objective values close to optimal. Lokhnygina and Tsiatis²¹ also used the backward induction algorithm to compute optimal two-stage designs for different optimality criteria and confirmed the finding by Jennison and Turnbull¹⁹ that optimal group-sequential designs are almost as efficient as optimal adaptive designs. Recently, Pilz et al²² applied the Euler-Lagrange equations to provide a pure variational solution of the problem to minimize the expected sample size under the alternative under constraints on maximum type I error rate and minimal power.

In parallel, optimal adaptive designs were developed for clinical trials with binary responses, that is, discrete outcomes, as well. Simon²³ proposed optimal two-stage group-sequential designs which either minimize the expected sample size under the null hypothesis or the maximum sample size. Among others, Jung et al²⁴ and Mander and Thompson²⁵ analyzed optimal group-sequential designs. Shuster²⁶ proposed an adaptive design that minimizes the maximum expected sample size under different effect sizes for a one-arm trial with binary outcome. This design was applied to continuous endpoints by Wason and Mander²⁷ and to the multi-stage case by Wason et al.²⁸ Banerjee and Tsiatis²⁹ applied the framework of Bayesian decision theory to derive two-stage designs with adaptive second stage which minimize the expected sample size under the null hypothesis. Their solution strategy was improved by Englert and Kieser³⁰ and by Kunzmann and Kieser^{31,32} who provided solutions for almost arbitrary objective criteria by applying integer linear programming to derive optimal designs.

A general solution strategy for the problem (1)-(3) for asymptotically normally distributed outcomes is implemented in the R³³-package `adopr`.³⁴ There, the variational problem is made finite-dimensional by discretizing the functions inside

the continuation region. This software package is used for all computations performed for the following examples. A brief description of how the optimization problem (1) - (3) is solved numerically in `adoptr` is provided in the Appendix.

3 | ILLUSTRATING CLINICAL TRIAL EXAMPLE

Currently, partial pancreateoduodenectomy (PD) is the indicated surgical procedure for a wide range of benign and malignant diseases and offers the only potential cure for pancreatic head cancer. The current gold standard, open PD (OPD), is performed via laparotomy. This procedure is associated with a substantial morbidity of approximately 40%, even in specialized centers.³⁵ For robotic PD (RPD), the surgeon operates a surgical robot facilitating increased dexterity, visualization, and range of motion as compared to laparoscopic PD. Thus, RPD might offer a viable alternative to OPD. However, a thorough investigation comparing RPD with the current gold standard OPD is lacking.

To fill this gap, a randomized controlled trial to compare RPD and OPD was proposed in a recently accepted grant proposal.³⁶ The primary outcome is morbidity within 30 days after surgery. It is measured by the comprehensive complication index (CCI).³⁷ The CCI was developed by surgeons and patients and considers the patient's perspective as well as objective parameters of surgical effectiveness. It ranges from 0 to 100 and can be assumed to be normally distributed³⁷ with variance $\sigma^2 = 400$.³⁸ A clinically relevant mean difference for the CCI is $\theta = 10$.³⁸

The intended clinical trial is of exploratory character and, therefore, no formal sample size calculation has been performed. We assume, however, that this trial is successful and results in a subsequent confirmatory trial that should be planned with a two-stage design. Note that the standardized clinically relevant effect size for the CCI is $\delta = \theta/\sigma = 0.5$. Following the ICH E9 guideline,³⁹ the one-sided type I error rate is to be strictly controlled at $\alpha = 0.025$. A minimal power at the effect size $\delta = 0.5$ of $1 - \beta = 0.9$ is required.

A standard approach to define a two-stage design in this setting is the choice of a group-sequential design with equally-sized stages. Such a design, labeled as D_1 , that controls the maximum type I error rate by the rule of Pocock⁸ can be obtained by the R-package `rpact`.⁴⁰ It is depicted in Figure 1 and its characteristics are specified in Table 1. The early-stopping boundaries are $c_f = 0.5$ and $c_e = 2.17$, and the stage-wise per-group sample size equals $n_1 = n_2 = 48$. The design shows a maximum sample size of 96 that is not much larger than the fixed design's sample size of 85 which would be necessary to fulfill the constraints on type I error and power in a one-stage design. The expected sample size under the alternative equals 65.5 patients per group and is, therefore, considerably lower than the single-stage design's sample size. The expected sample size under the null hypothesis amounts to 62.1 per-group patients.

While the group-sequential design shows a satisfactory performance with respect to maximum sample size and expected sample size under the alternative, its conditional power to show an effect of $\delta = 0.5$ drops to a value of 45.3% at the early-futility boundary $c_f = 0.5$, that is, for the smallest value of z_1 that implies continuation of the trial. Note that from a purely statistical perspective, a further restriction of the design would deteriorate its performance with respect to the underlying unconditional objective criterion. In practice, however, two-stage designs with a too low conditional power are often viewed critically and difficult to communicate. Conducting a second stage with a very small (or large) conditional power would imply that the result of the trial would already be predictable at the interim analysis with a high probability. However, recruiting additional patients may only be worth the effort if there is a sufficient chance that \mathcal{H}_0 can be rejected at the final analysis. Consequently, a pre-planned design with a too low conditional power may cause a design modification at the interim analysis when unblinded trial data becomes available. Such a design reassessment is discouraged by regulatory authorities.⁴¹ Furthermore, it implies a performance deterioration since the desire of a sufficiently large conditional power could have been considered during the planning stage.

To overcome this drawback, it is a common approach to modify the design's second-stage sample size such that the conditional power in case of entering stage two should be at least above a certain boundary that is chosen to equal $1 - \beta_c = 0.7$ in this example. However, this is a user-specific threshold which might be chosen even smaller. To ensure a conditional power of 70% for any possible interim outcome inside the continuation region, the sample size is modified according to Formula (4) if the conditional power target value cannot be reached with the group-sequential design. Otherwise, if the conditional power of the group-sequential design equals already at least 70%, the trial continues with the initially planned stage-two sample size of the group-sequential design.¹² The resulting design D_2 is depicted in Figure 1 and its key characteristics are summarized in Table 1. All first-stage parameters are equal to the underlying group-sequential design by definition. The sample size function is monotonically decreasing and convex. Note that for large values of z_1 the stage-two sample size equals the initially planned $n_2 = n_1 = 48$ patients per group. Furthermore, the conditional power constraint is indeed fulfilled within the entire continuation region. While a fixed design would require 85 patients per

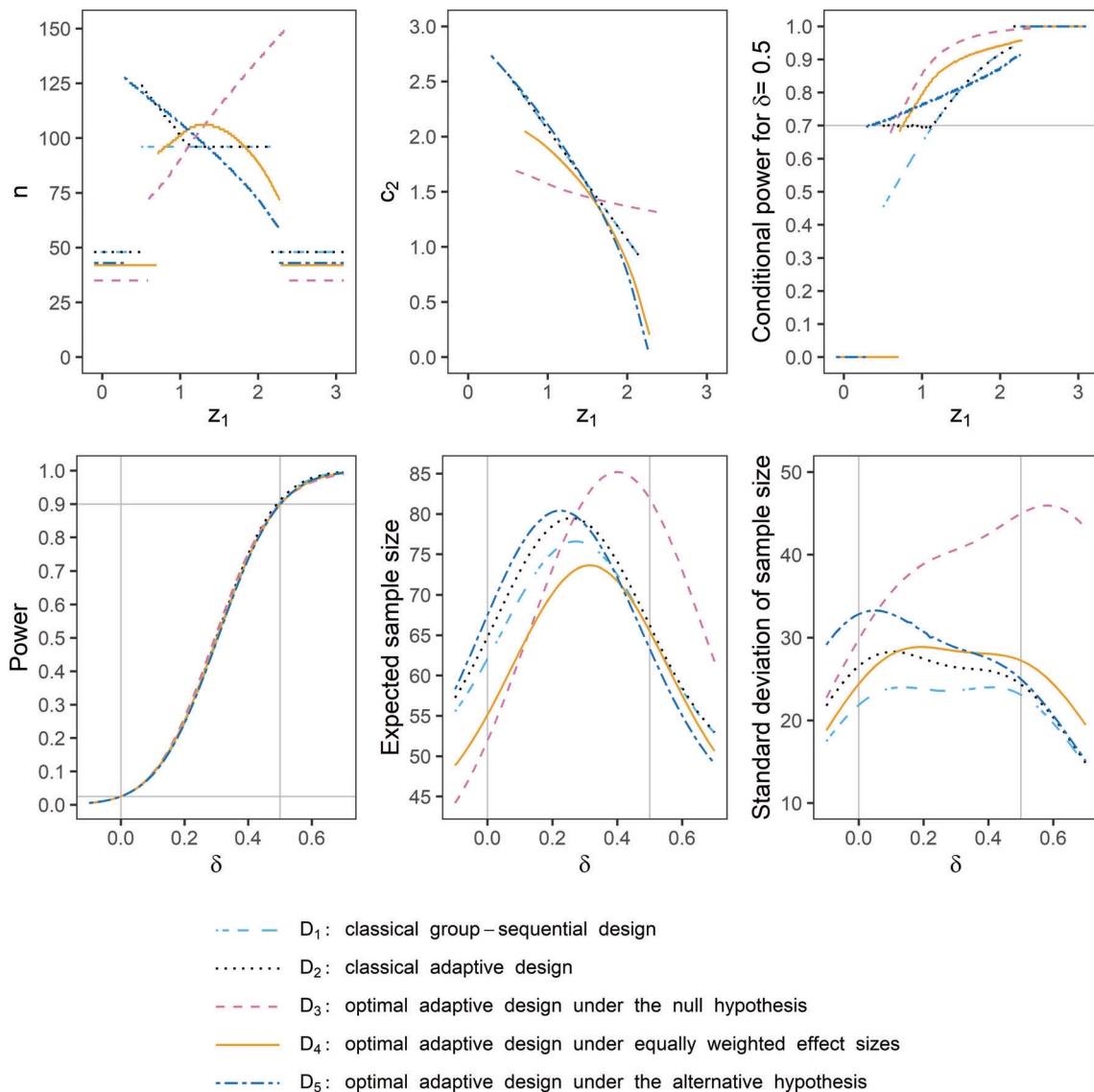


FIGURE 1 Characteristics of different adaptive two-stage designs that control the type I error rate at 2.5% and achieve a power of 90%. Apart from the group-sequential design, all designs require a conditional power of at least 70% in case of entering the second stage [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Characteristics of different adaptive two-stage designs

Design	Introduced in	n_1	c_f	c_e	$E_{\delta=0.5}[n]$	$E_{\delta=0}[n]$	$\max_{z_1}\{n_1 + n_2(z_1)\}$
D_1	Section 3	48	0.50	2.17	65.5	62.1	96
D_2	Section 3	48	0.50	2.17	66.2	64.8	124
D_3	Section 4	36	0.59	2.40	81.9	52.0	152
D_4	Section 4	43	0.71	2.29	65.5	55.1	106
D_5	Section 4	44	0.28	2.27	63.3	67.5	128
D_6	Section 6.1	44	0.28	2.27	63.3	67.5	127
D_7	Section 6.2	50	0.10	2.40	70.8	116.6	334
D_8	Section 6.3	44	0.00	2.27	63.5	70.1	107
D_9	Section 6.4	55	0.64	2.31	66.0	64.8	98

group, the expected sample size of this adaptive design under the point alternative $\delta = 0.5$ equals 66.2 and the expected sample size under the null hypothesis $\delta = 0$ equals 64.8. In contrast to the group-sequential design, the maximum sample size of this design increases to a large value of 124 subjects per group. Note that the power of the modified design equals 91.1% and thus the design is slightly over-powered. In practical application, the design's overall power is often not taken into account when the design's sample size is recalculated. Since the power of a clinical trial design is highly relevant for multi-stage designs as well,²⁰ it should, however, be considered as a decision tool for design planning.

4 | BENEFIT OF OPTIMIZING TWO-STAGE DESIGNS

The adaptive design D_2 satisfies all desired constraints outlined earlier but its parameters were fixed heuristically independent of a well-defined objective criterion. Instead of pre-planning a group-sequential design that fulfills a condition on unconditional power and recalculating the sample size to ensure a certain conditional power, one can define an optimal design that fulfills both constraints simultaneously. This approach takes the unconditional power as well as the conditional power into account and thus, neither the entire trial nor its second stage (in case of entering it) show an unacceptable probability to reject the null hypothesis if the alternative is true. Furthermore, due to the optimality of the resulting design, it is more efficient than any heuristic procedure. Note that, as outlined above, a constraint on conditional power may increase the acceptance of a two-stage design in practice since there is no risk to end up with an under-powered second stage.

When an optimal design is to be defined, the choice of the objective criterion is of high importance and not straightforward.⁴² Barber and Jennison¹⁵ proposed to minimize a weighted sum of expected sample size under the null and under the alternative hypothesis. This can be interpreted as a point prior on the two effect sizes $\delta = 0$ and $\delta = 0.5$ where the probability $p \in [0, 1]$ is assigned to the alternative hypothesis and the probability $1 - p$ to the null hypothesis. Plugging this together with the required constraints on the design's operating characteristics in (1)-(3) yields the optimization problem

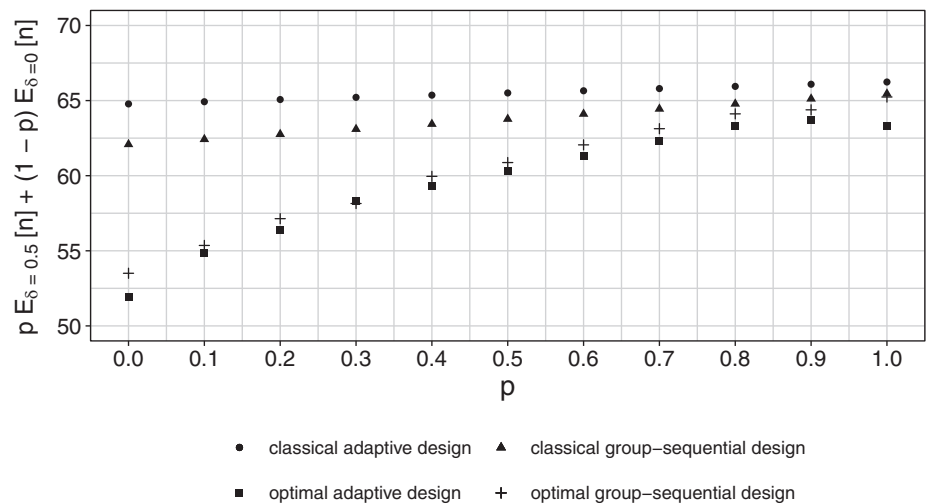
$$\begin{aligned} & \underset{D}{\text{minimize}} && p E_{\delta=0.5}[n] + (1 - p) E_{\delta=0}[n] \\ & \text{subject to} && \text{Type I error rate} && \leq 2.5\% \\ & && \text{Power at } \delta = 0.5 && \geq 90\% \\ & && \text{Conditional power at } \delta = 0.5 && \geq 70\%. \end{aligned}$$

We solve this problem for three values of p :

- $p = 0$ corresponding to minimization of expected sample size under the null hypothesis; this results in design D_3 ,
- $p = 0.5$ corresponding to equal weighting of the null and the alternative hypothesis; this results in design D_4 ,
- $p = 1$ corresponding to minimization of expected sample size under the alternative hypothesis; this results in design D_5 .

The resulting designs are plotted in Figure 1 and their main characteristics are reported in Table 1. The features of the optimal designs are highly dependent on the underlying objective criterion. In case of $p = 0$, the sample size function is monotonically increasing. This property is intuitive since lower sample sizes for small values of z_1 imply better performance with respect to the objective criterion "expected sample size under the null hypothesis." Larger values of z_1 are quite unlikely under \mathcal{H}_0 and for this reason, larger sample sizes for these values do hardly affect the objective criterion while they compensate for the smaller sample sizes for z_1 close to c_f to fulfill the power constraint. The continuation region of design D_3 is $[c_f, c_e] = [0.59, 2.4]$. The maximum sample size of the optimal design D_3 equals 152 and is hence much larger than the fixed design's sample size of 85. The critical value function $c_2(\cdot)$ is very flat in this setting. When the expected sample size under the alternative is part of the objective function, the critical value function becomes concave. For optimization exclusively under the alternative, the sample size function is monotonically decreasing with a maximum sample size of 128. If the null and the alternative hypothesis are weighted equally, the sample size function is almost constant for small effect sizes and decreasing for larger effect sizes. The regularizing aspect of weighting two effect sizes against each other is clearly visible by the lower maximum sample size of 106. However, this value is still larger than the maximum sample size of 96 of the classical group-sequential design from the previous section. The enormously large maximum sample sizes of the designs D_3 and D_5 make these designs unattractive in practice due to the risk to conduct

FIGURE 2 Comparison of achieved values of objective functions for optimal and non-optimal designs. All designs are chosen to control the type I error rate at 2.5% and achieve a power of 90%. The optimized designs minimize the respective objective criterion whose value is plotted on the y-axis and, furthermore, require a conditional power level of 70%



an enormously large trial. Because of its high relevance, the maximum sample size will become part of the optimization problem in Section 6.3.

The conditional power is strictly increasing with z_1 for the three optimized designs D_3 , D_4 , and D_5 and is for all designs uniformly larger than the conditional power of design D_2 within the entire continuation region. This inferiority in terms of conditional power implies that D_2 needs a larger stage-one sample size of $n_1 = 48$ compared with $n_1 = 36$ (D_3), $n_1 = 43$ (D_4), and $n_1 = 44$ (D_5) to achieve the overall power condition. We would like to highlight the interesting observation that the presented designs exhibit highly different characteristics even though they are restricted by the same constraints and, therefore, show almost identical power curves.

To allow efficiency comparison, the optimal design was computed for further values of $p \in [0, 1]$ and the differences between the optimal value of the objective function and the value achieved by the classical designs D_1 and D_2 were calculated. The achieved values of the objective function in dependence of p for the respective optimal design and the classical designs D_1 and D_2 are illustrated in Figure 2. The performance of D_1 and D_2 deteriorates with increasing weight on the null hypothesis, that is, lower p . When the objective function is “expected sample size under the null hypothesis” ($p = 0$), the classical adaptive design D_2 shows an expected sample size of 64.8 while an optimal value of 51.9 with design D_3 is possible. The classical group-sequential design D_1 achieves a value of 62.1. Note, however, that this design is not restricted by a conditional power constraint as the other designs are. With increasing p , the objective function values of the conventional designs D_1 and D_2 become closer to the optimal one. The performance gap is smallest when minimizing solely under the alternative ($p = 1$) where D_1 and D_2 show an expected sample size of 65.5 and 66.2, respectively, compared to an optimal value of 63.3. This observation allows the conclusion that choosing an adaptive two-stage design based on conditional power considerations is not far from being optimal if the focus is on the expected sample size under the alternative. However, if others than this single effect size are under consideration, more efficient design choices are available.

5 | OPTIMAL GROUP-SEQUENTIAL DESIGNS

Group-sequential designs, that is, designs with a constant $n_2(\cdot)$ -function, are a special case of adaptive designs. The question of the efficiency of group-sequential designs in comparison to generic two-stage designs has been addressed in various publications.^{19,43,44} Applying a group-sequential design simplifies the interim analysis to a “go/no-go” decision where it is only decided whether or not the second stage is conducted; the stage-two sample size does not further depend on the interim result. In Section 3, it was shown that a conventional group-sequential design shows an acceptable performance in terms of “expected sample size under the alternative” and a low maximum sample size that is a highly desirable property. However, its performance can be improved by an optimal adaptive design if the expected sample size under the null hypothesis becomes a noticeable part of the objective criterion. Furthermore, the classical group-sequential design drastically misses the conditional power target value of 70% what may be a drawback of this design. Requiring a constraint on conditional power in a group-sequential design may lead to a remarkably efficiency reduction due to the restriction to

TABLE 2 Stage-wise sample sizes and early-stopping boundaries of optimal group-sequential designs that fulfill a conditional power constraint

p	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
n_1	37	38	39	40	42	43	44	45	46	46	47
n_2	69	65	62	60	58	56	55	54	53	52	51
n_2/n_1	1.86	1.71	1.59	1.50	1.38	1.30	1.25	1.20	1.15	1.13	1.09
$n_1 + n_2$	106	103	101	100	100	99	99	99	99	98	98
c_f	0.68	0.68	0.69	0.70	0.72	0.72	0.72	0.72	0.71	0.70	0.68
c_e	2.40	2.40	2.39	2.32	2.27	2.24	2.22	2.20	2.19	2.18	2.17

constant $n_2(\cdot)$ -functions. The conditional power is monotonically increasing in both n_2 and z_1 since more evidence against the null hypothesis has already been collected if z_1 is large. A generic two-stage design allows modifying the sample size such that a conditional power constraint can be fulfilled within the entire continuation region (cf. Equation (4)). This implies that a fixed conditional power constraint requires lower sample sizes for larger effect sizes observed at interim. In a group-sequential design, however, the constant sample size n_2 has to be chosen such that the conditional power constraint is fulfilled within the entire continuation region. Therefore, one may expect a subpar performance of group-sequential designs when conditional constraints are imposed.

To investigate the efficiency of optimal group-sequential designs, we minimize the previous objective criterion consisting of a weighted sum of expected sample size under the null and the alternative hypothesis under constraints on type I error rate, overall power, and conditional power by solving

$$\begin{aligned}
 & \underset{D}{\text{minimize}} && p E_{\delta=0.5}[n] + (1-p) E_{\delta=0}[n] \\
 & \text{subject to} && \text{Type I error rate} && \leq 2.5\% \\
 & && \text{Power at } \delta = 0.5 && \geq 90\% \\
 & && \text{Conditional power at } \delta = 0.5 && \geq 70\% \\
 & && n_2(z_1) && \equiv n_2.
 \end{aligned}$$

The above optimization problem is solved for $p \in \{0, 0.1, \dots, 1\}$ and the corresponding optimal adaptive and optimal group-sequential designs are computed. In Figure 2, the value of the objective function is plotted against p for the optimal generic two-stage design and the optimal group-sequential design from the respective optimization problem. The performance differences appear to be quite small. The maximum difference occurs when $p = 1$ and equals to 1.9 patients per group. This difference is much smaller than the maximum difference between an optimized design and the two-stage design D_2 that was chosen without any optimality considerations. Interestingly, the additional restriction of the conditional power does not imply a large increase in maximum sample size. While the maximum sample size of the conventional group-sequential design D_1 equals 96, the maximum sample size of the optimized group-sequential designs varies between 98 and 106 (cf. Table 2). However, this implies that the early-futility-boundary of these optimized designs is increased in comparison with the approaches from Sections 3 and 4. This can be explained by the large sample sizes that would be necessary to ensure a conditional power of 70% for small values of z_1 . Consequently, an aggressive futility stop avoids these situations and leads to acceptable maximum sample sizes and conditional power values for all $z_1 \in [c_f, c_e]$ simultaneously.

These observations allow stating that the restriction to group-sequential designs does not cause notable efficiency reduction even if a constraint on the conditional power is imposed. It is important to note that the group-sequential designs were chosen to be fully optimal, in particular without any assumption on the ratio between the first- and second-stage sample sizes. In practical applications, however, group-sequential designs are often chosen with equally sized stages. Jennison and Turnbull¹⁹ observed that group-sequential designs with equally sized stages do not cause a large efficiency reduction if the critical values are chosen optimally. However, no conditional power constraint was incorporated in their framework. Since a constraint on conditional power implies that the stage-two sample size n_2 must be sufficiently large, allowing the ratio between n_1 and n_2 to vary may be useful to avoid over-powered studies. In Table 2, the first- and second-stage sample sizes of the optimal group-sequential designs for the example presented here are listed.

Interestingly, the optimal second-stage sample size is always larger than the first-stage sample size. The ratio between the sample sizes of the second and the first stage varies between 1.09 and 1.86. As outlined above, this may be caused by the conditional power constraint that requires sufficiently large stage-two sample sizes to guarantee a conditional power of 70%. As a further interesting observation, the continuation region of an optimal group-sequential design slightly shrinks with increasing p .

6 | CUSTOMIZATION OF OPTIMAL DESIGNS

In this section, it is described how an optimal two-stage design can be modified if it shows properties that may not be desired by the trial investigators. Different optimal adaptive designs will be presented that solve optimization problems guaranteeing specific characteristics. It is well-known that an optimized classical group-sequential design that tests the test statistic from all $n_1 + n_2$ patient data at the final analysis against a critical value c is a sufficient approximation of an optimal adaptive design.^{15,19,20} In this case, $c_2(\cdot)$ only depends on the critical value c of this test and is given by

$$c_2(z_1) = \sqrt{\frac{n_1 + n_2}{n_2}} \cdot c - \sqrt{\frac{n_1}{n_2}} \cdot z_1.$$

Therefore, such a design can be represented by the five values n_1 , c_f , c_e , n_2 , and c . To investigate whether this approximation is also valid for the broad variety of optimization problems that are solved in the following, the corresponding optimal five-parametric group-sequential design and the achieved objective value are reported in each subsection as well. Note the difference to Section 5 where arbitrary critical value functions $c_2(\cdot)$ were allowed. This increased flexibility enabled the group-sequential design to fulfill a constraint on conditional power. In the simplified setting of a five-parametric group-sequential design, such a conditional constraint that must be fulfilled for each $z_1 \in [c_f, c_e]$ separately cannot be guaranteed in general.

6.1 | Typical optimization problem

We still consider the pancreatic surgery trial from Section 3 and start planning an adaptive two-stage design based on the typical optimization problem to minimize “expected sample size under the alternative” under constraints on maximum type I error rate and minimal power. Therefore, the optimization problem to solve is

$$\begin{array}{lll} \underset{D}{\text{minimize}} & E_{\delta=0.5}[n] & \\ \text{subject to} & \text{Type I error rate} & \leq 2.5\% \\ & \text{Power at } \delta = 0.5 & \geq 90\%. \end{array}$$

In Figure 3, the resulting design D_6 is illustrated. The stage-two sample size function as well as the stage-two critical value function are monotonically decreasing and concave. The maximum sample size of 127 appears at the boundary for an early futility stop in the continuation region $[c_f, c_e] = [0.28, 2.27]$. Since the minimal level of the conditional power is 69%, the conditional power constraint in the previous examples was indeed necessary to achieve a conditional power of at least 70%. The expected sample size under the alternative amounts to 63.3. Furthermore, the constraints on type I error rate and overall power are met precisely. The properties of the optimal five-parametric group-sequential design that is solving this optimization problem are summarized in Table 3. The performance gap with respect to the expected sample size under the alternative is quite small and amounts to 3.3 patients per group.

6.2 | Modification of the power curve

Intuitively, there are situations where the trial sponsor may not agree with constraints imposed solely on power and type I error rate. In contrast, there may be specific further requirements. We already showed in Section 4 that a constraint on

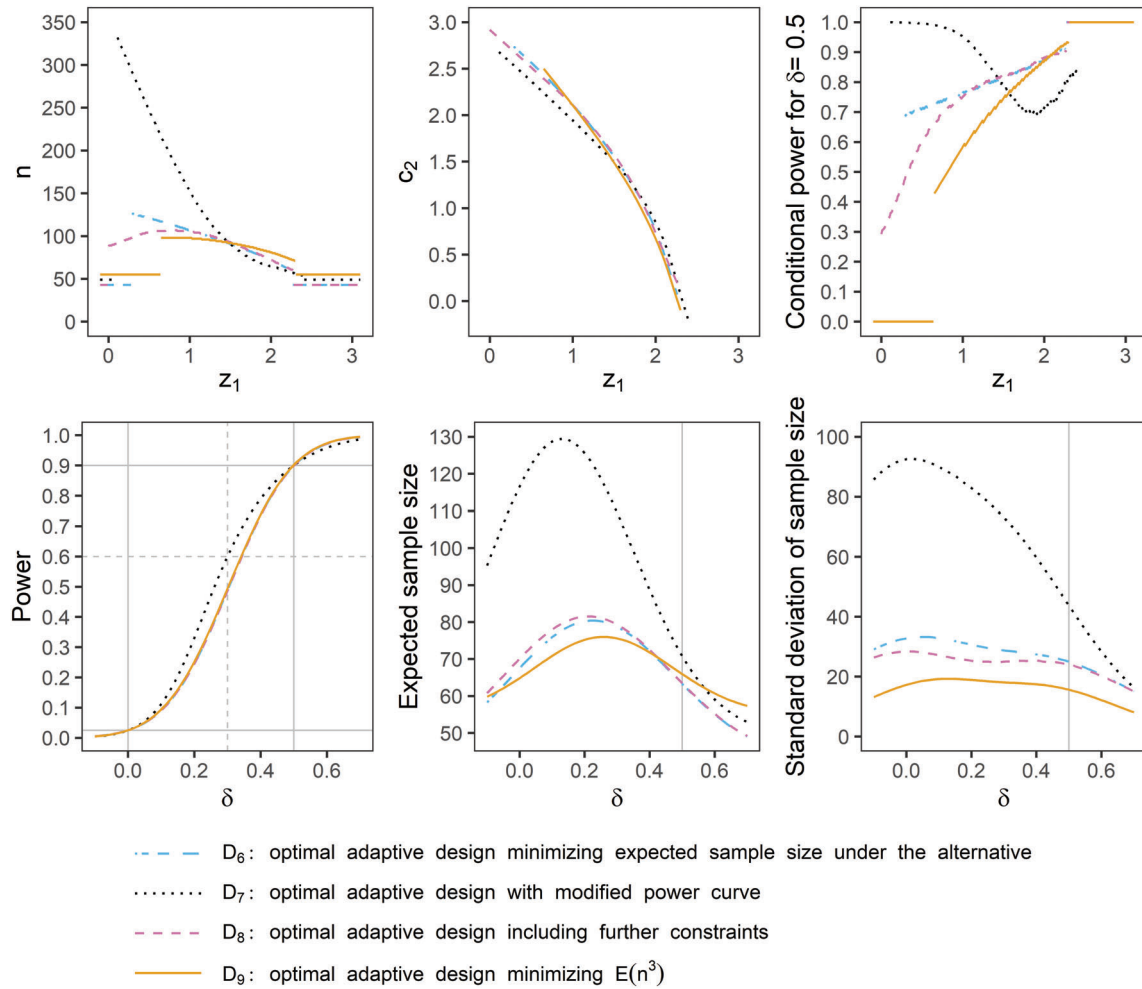


FIGURE 3 Characteristics of optimal two-stage designs for different optimization problems. The concrete definitions of the underlying optimization problems are given in the corresponding Sections 6.2 to 6.4 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Characteristics of optimal group-sequential designs based on five parameters and comparison to optimal adaptive designs

Introduced in	n_1	n_2	c_f	c_e	c	Objective value	Value of optimal adaptive design	Efficiency reduction
Section 6.1	47	64	0.81	2.06	2.33	66.57	63.31	5.1%
Section 6.2	54	83	0.65	2.14	2.19	78.44	70.83	10.7%
Section 6.3	46	49	-0.03	2.15	2.21	64.77	63.52	2.3%
Section 6.4	57	37	0.92	2.18	2.10	344 068.2	338 583.5	4.3%

conditional power can be incorporated in the optimization problem. Besides conditional power, there may be further possible considerations. Mehta and Patel⁴⁵ pointed out that an advantage of adaptive designs may be that anticipated power values can be guaranteed not only for one particular alternative effect size. While fixed designs only allow fixing two points on the power curve, it is possible to define more than two power values within an adaptive design. Imagine in the surgery example that a mean difference of $\theta = 6$ in CCI, corresponding to a standardized effect size of $\delta = 0.3$, is still of clinical relevance. The optimal design from Section 6.1 shows a power of 48.7% to detect this effect size and is, therefore, comparable with a coin flip. However, it may be desirable to increase the power at $\delta = 0.3$ to 60%. Adding this requirement to the previous optimization problem yields

minimize	$E_{\delta=0.5}[n]$		
subject to	Type I error rate	\leq	2.5%
	Power at $\delta = 0.5$	\geq	90%
	Power at $\delta = 0.3$	\geq	60%.

Figure 3 shows the resulting design D_7 . The sample size function is monotonically decreasing as well as the critical value function, which is hardly changed in comparison with D_6 . The price of the second power constraint measured in maximum sample size is huge. While design D_6 shows a maximum sample size of 127 per group, the maximum sample size of design D_7 amounts to 334 patients per group. With regard to the objective criterion ϵ expected sample size under the alternative, the performance loss is not that drastically. While D_6 achieved an objective function value of 63.3, the design with a further power constraint shows an expected sample size at $\delta = 0.5$ of 70.8. For all effect sizes $\delta \in (0, 0.5)$, the difference between the expected sample size curves is, however, much larger. An optimal group-sequential design that fulfills both power constraints shows an expected sample size at $\delta = 0.5$ of 78.4. The conditional power of design D_7 to detect the effect size $\delta = 0.5$ is very large within the continuation region since larger sample sizes are necessary to obtain an overall power of 60% for the effect size $\delta = 0.3$. The enormous increase in maximum sample size implies that the standard deviation of the sample size is increased. For instance, the standard deviation under the alternative equals 43.7 compared with a value of 25 for the previous design D_6 .

6.3 | Inclusion of further constraints

The example in Section 6.2 illustrated that even though an adaptive design allows the incorporation of multiple power constraints, this comes with a massive increase in maximum sample size. However, the maximum sample size a design may induce is an important feature for the trial investigators. If there are scenarios where the sample size of a two-stage design is much larger than the fixed design's sample size, investors may not be willing to conduct the trial with an adaptive design due to the risk to end up with a too large trial duration. The results by Jennison and Turnbull¹⁰ and Barber and Jennison¹⁵ show that for group-sequential designs a moderate decrease in maximum sample size above the fixed design's sample size is sufficient to obtain a substantial efficiency gain. Restricting the maximum sample size of an adaptive design to a specific multiple of the fixed design's sample size may therefore help to obtain designs that are still efficient but more appealing from an investor's view. In our example, we assume that the trial sponsor is not willing to invest more than 1.25 times the fixed design's sample size of $n = 85$. This restriction is included as a further constraint. Another issue that is often discussed in the planning stage of an adaptive design is the choice of the futility stop boundary. The optimal design D_6 from Section 6.1 shows an early-futility boundary of $c_f = 0.28$. However, stopping a clinical trial early for futility even if the interim effect points in the desired direction (ie, $c_f > 0$) might be unattractive for an investigator. In order to avoid situations where a clinical trial with a positive interim effect is stopped early for futility, a constraint $c_f \leq 0$ is required in our example. This constraint as well as the constraint on the maximum sample size can be included in the optimization framework, and the minimization problem to solve becomes

minimize	$E_{\delta=0.5}[n]$		
subject to	Type I error rate	\leq	2.5%
	Power at $\delta = 0.5$	\geq	90%
	$\max_{z_1} \{n_1 + n_2(z_1)\}$	\leq	107
	c_f	\leq	0.

The obtained design D_8 , which is depicted in Figure 3, is then the optimal two-stage design with respect to the expected sample size under the alternative $\delta = 0.5$ under all possible design choices that fulfill the above collection of constraints. The continuation region now is $[c_f, c_e] = [0, 2.27]$. Hence, the optimal futility-stop boundary takes its maximum value of $c_f = 0$, and the efficacy-stop boundary has hardly changed in comparison with D_5 . The sample size function is no longer strictly monotonically decreasing but quite flat for values of z_1 close to c_f . This is necessary to not exceed the maximum sample size of 107. Conversely, the stage-two critical value function is still monotonically decreasing and concave. In contrast to the examples in Sections 4 and 5, the resulting design shows an acceptable maximum sample size and a low

probability of an early futility stop. However, this comes with the cost of a decreased conditional power that equals 29.2% at the early-futility boundary. While the inclusion of further constraints changes the design's properties by fitting them to the trial sponsor's preferences, this implies an increased value of the objective function. Indeed, one can exactly state the price of the included constraints in terms of "expected sample size under the alternative." The expected sample size under the alternative $\delta_1 = 0.5$ equals 63.5 subjects per group for this stronger restricted design D_8 compared to 63.3 for the optimal design D_6 without those additional constraints. Therefore, the inclusion of these constraints causes an efficiency loss of 0.2 expected subjects per group under the alternative. Interestingly, this difference appears to be very small. This observation indicates that adding additional operational constraints to the basic optimization problem may be not very expensive in terms of the objective criterion while on the other side making the design much more suitable for clinical application. If the design is restricted to a group-sequential design based on five parameters, the expected sample size under the alternative slightly increases to a value of 64.8 patients per group.

6.4 | Customization of the objective function

A further possibility of adjusting an optimal two-stage design is the utilization of another objective function. From an applicant's view, the expected sample size under one or several effect sizes may be the most important objective to minimize. In order to flatten the sample size curve, one may, however, consider the minimization of a higher moment of the sample size function $n(z_1) = n_1 + n_2(z_1)$. Minimizing a higher than the first moment of n still leads to small expected sample sizes but, in addition, larger sample sizes are penalized stronger. This should imply that the sample size curve is flattened what may be preferred in practice. For our example, we assume that the third moment under the alternative, that is $E_{\delta=0.5} [n^3]$, is to be minimized. Note that this is an arbitrary choice and minimizing higher or lower moments would lead to more or less flattened sample size functions, respectively. To minimize the third moment of n under constraints on type I error rate and power, the optimization problem to solve is:

$$\begin{array}{lll} \underset{D}{\text{minimize}} & E_{\delta=0.5} [n^3] & \\ \text{subject to} & \text{Type I error rate} & \leq 2.5\% \\ & \text{Power at } \delta = 0.5 & \geq 90\%. \end{array}$$

The resulting optimal design D_9 is plotted in Figure 3. The sample size function is monotonically decreasing but much flatter than for design D_6 . In particular, the first-stage sample size is increased to $n_1 = 55$. Therefore, smaller stage-two sample sizes are sufficient to fulfill the power condition. This implies that the maximum sample size decreases to 98 and the standard deviation under the alternative decreases to 15.6. The critical value function $c_2(\cdot)$ has hardly changed and is monotonically decreasing and concave. The continuation region becomes $[c_f, c_e] = [0.64, 2.31]$ and is thus slightly shifted to the right compared with design D_6 . The minimal conditional power in case of continuation has a disappointing value of 42.3% at the early futility stop boundary for design D_9 . The expected sample size under the point effect size $\delta_1 = 0.5$ equals 66 and is therefore slightly larger than for design D_6 where a value of 63.3 was achieved. A group-sequential design with five parameters that solves the same optimization problem shows an expected sample size at $\delta = 0.5$ of 66.1. In terms of the objective criterion "third moment of the sample size under the alternative," the simpler group-sequential design is 4.3% less efficient than an optimal adaptive design (cf. Table 3). In total, this example illustrates that flatter sample size curves can be obtained by employing another objective function than "expected sample size." While this is of high theoretical appeal, this objective criterion may be difficult to interpret and communicate in practical application. Therefore, imposing an upper bound to the maximum sample size may be better suited in a concrete clinical trial. However, it should be mentioned that choosing another objective function instead of adding an additional constraint does not additionally reduce the solution space of the optimization problem. Therefore, one may not end up with an optimization problem that is unsolvable if the objective function is changed while constraints on, for example, conditional power and maximum sample size might be contradictory.

7 | DISCUSSION

In this article, we analyzed different aspects of the optimal choice of a two-stage design. When it is decided to specify the entire design during the planning stage, the task of choosing a suitable design can be embedded in an optimization

problem that leads to the best possible design for the selected objective criterion under specified constraints. We believe that applying these designs or optimal group-sequential designs that are valid approximations of optimal adaptive designs is an appealing option for clinical research. In the presented examples, we demonstrated that the optimal design's characteristics strongly depend on the underlying optimization problem. In particular, an adaptive design based on conventional rules can be inefficient if the expected sample size under the alternative is not the only objective criterion (cf. Section 4). Consequently, there is no "one-size-fits-all approach" to define a generic two-stage design. Instead, the trial design (ie, the design parameters) should be chosen case-driven, ideally in an optimal way. Studying the optimal design for a particular situation can also be helpful to detect such situations where an originally chosen design performs inefficiently. By comparing the classical design with the optimal one, the latter can serve as benchmark that allows quantifying the performance loss of the non-optimal design. Similarly, optimal designs can be used to quantify the cost of additional constraints. Deriving the optimal designs with and without the constraint in question allows to state the efficiency loss incurred by the inclusion of the constraint in terms of the chosen objective criterion. In the examples considered in this manuscript, we did not observe a substantial performance decline after adding further operational constraints (cf. Section 6.3), or restricting to group-sequential designs (cf. Section 5) as long as the power curve remained identically. This shows that the reduction in expected sample size of two-stage as compared to fixed designs mainly stems from the possibility of early stopping and not from the shape of the stage-two functions. Modifying the power curve results in substantial additional costs (cf. Section 6.2).

In this manuscript, a variety of adaptive and group-sequential designs was presented. All of them showed different characteristics that were fitting to particular optimization problems. The choice of the objective criterion is not straightforward. Indeed, different objective criteria lead to very different designs that may show unfavorable properties. Therefore, defining an optimization problem that takes different performance criteria into account may be a reasonable procedure. Note that the optimization approach is naturally limited by the solution space that is spanned by its constraints. In Section 3, a conventional group-sequential design was introduced that was not derived by any optimization approach. While this design showed a low maximum sample size, its conditional power was low for small values of z_1 inside the continuation region. Including a conditional power constraint (Section 5) led to increased conditional power values without enormously exceeding the previous maximum sample size (cf. Table 2). However, those optimized design showed a quite aggressive early futility stop. This observation is as expected since small values of z_1 inside $[c_f, c_e]$ require larger stage-two sample sizes n_2 to exceed a certain conditional power level. Consequently, a design with a liberal futility stop, a low maximum sample size, and a large conditional power for any value $z_1 \in [c_f, c_e]$ does not exist. Therefore, each time when an adaptive (or a group-sequential) design is planned by optimization, one has to choose the underlying constraints very carefully and weigh them against each other with caution. It depends on the concrete situation which of the large number of desired properties is dispensable in the specific clinical trial.

In particular the conditional power often plays a pivotal role in adaptive sample size adjustments during unblinded interim analyses (cf. for instance Chapter 7 of the book by Wassmer and Brannath²). Instead of raising the sample size post-hoc during an interim analysis, one can choose an optimal design under a constraint on the minimal conditional power. If the design is pre-planned optimally under a conditional power constraint, a foreseeable mid-course design reassessment can be avoided. This is recommended by regulatory authorities.⁴¹ Of course, adaptive designs still enjoy the feature of the possibility to modify the trial design at the interim analysis. A constraint on conditional power, however, can already be formulated and ensured during the planning phase as long as the planning assumptions do not change. Furthermore, the optimization approach implies that the corresponding design shows the best possible performance under the given objective criterion and constraints. Any design modification during the ongoing trial would lead to a performance deterioration. An optimal design with a constraint on minimal conditional power thus joins both the conditional and the unconditional perspectives on design performance. While a conditional power constraint is not required by regulatory authorities, it can effectively prevent rare situations in which conditional power would drop so low that a sponsor might again be tempted to conduct an ineffective sample size recalculation. The small costs in terms of (unconditional) expected sample size are outweighed by the additional intuitive appeal and the protection against a potentially perceived need to conduct an ineffective post-hoc sample size recalculation.

Besides the conditional power, the futility stop is another important characteristic of a two-stage design. It must be noted that the futility stop is necessarily binding in the setting of this article. An entirely non-binding futility stop cannot be provided in the outlined setting because the stage-two functions need to be defined on a compactum. This fact might be problematic in a phase III confirmatory trial where one might wish to consider many other factors, such as safety data or results from secondary endpoints, before terminating a trial for futility. In Section 6.3, it has been shown that

an upper-bound on the early-futility-stop boundary can be included in the optimization procedure. This seems to be the most suitable option in practice since this upper boundary can be chosen very small. Therefore, the choice of a strict upper bound on the futility stop may be acceptable in practice since entering the second stage with an interim effect that drastically points in the wrong direction is highly questionable from an ethical perspective.

The presented approach to optimal planning of two-stage designs can be restricted to group-sequential designs (cf. Sections 5 and 6). Those designs with constant sample size functions are a special case of generic two-stage designs. With regards to design performance, restricting the stage-two sample size to be constant appears to cause only a marginal difference in terms of efficiency. This result corresponds with the observations of Jennison and Turnbull,¹⁹ Lokhnygina and Tsiatis,²¹ and Posch et al.⁴⁶ Note that for the optimal designs considered in this manuscript, the design modifications are already outlined hypothetically for all possible interim outcomes during the planning stage. From this perspective, there is no need to choose a constant stage-two sample size function. Furthermore, a group-sequential design does not necessarily show a lower variability in terms of sample size than an adaptive design. If the first-stage sample size of a group-sequential design is quite small and a constraint on overall power is included, there may be a large jump in the design's sample size function (cf. Table 2). This makes the final sample size of the trial hardly predictable and hinders reliable planning of trial resources. Less variability in the trial's sample size can be achieved by, for example, including an upper bound on the maximum sample size (cf. Section 6.3). Nonetheless, the restriction to group-sequential designs provides at least three interesting features. Firstly, there are merely two possible sample sizes that can occur: the first-stage sample size n_1 if the trial is stopped early or the total sample size $n_1 + n_2$ if the trial is continued at the interim analysis. This may simplify the organization of the trial because only two potential scenarios have to be prepared. Secondly, group-sequential designs allow the usage of error-spending tests.^{47,48} Those can handle unpredictable variations in the actual group sizes. Barber and Jennison¹⁵ show that there are families of error-spending tests that are close to optimal for various objective functions. Hence, applying an efficient error-spending test and an efficient stage-wise sample size allocation in the group-sequential design may cause only small performance deterioration while simultaneously providing the option to react on unforeseen group sizes in the actual analysis. Finally, optimizing a simplified version of $c_2(\cdot)$ in the group-sequential framework may help to reduce the complexity of the optimization problem. It is computationally much easier to obtain a single optimal stage-two sample size n_2^* instead of a function $n_2^*(\cdot)$. Since the presented optimal stage-two critical value functions are not far from being linear, a sufficient approximation of optimal two-stage designs by constant n_2 - and linear c_2 -functions is an option worth to be considered. In particular, it is known from the literature that optimized classical group-sequential designs that test the test statistic from all $n_1 + n_2$ patient data at the final analysis against a critical value c are sufficient approximations of optimal adaptive designs.^{15,19,20} In this case, $c_2(\cdot)$ only depends on the critical value c of this test. In the examples of Section 6, it has been confirmed that such designs are a valid approximation of optimal adaptive designs with arbitrary $n_2(\cdot)$ - and $c_2(\cdot)$ -functions for a broad variety of optimization problems. This reduces the parameter space of the optimization problem to five parameters (n_1 , c_f , c_e , n_2 , and c). Investigating the computation of optimal designs for multi-arm multi-stage trials under these simplified assumptions may be an interesting topic for future research.

Planning based on a single-point alternative hypothesis $\delta = \delta_1$ is quite fragile to misspecification of δ_1 . We presented methods to regularize the optimization problem by imposing constraints (cf. Section 6.3) or customizing the objective function (cf. Section 6.4). Both ideas may increase the design's usability by providing desirable properties but come at the cost of a decreased efficiency in terms of the original objective criterion. We further regularized the design by not only minimizing under the point alternative hypothesis but under a weighted sum of null and alternative hypothesis (cf. Section 4). These considerations may be extended to a Bayesian approach, which allows becoming more robust against wrong planning assumptions on δ_1 . Instead of a point effect size, one may choose a prior distribution $\delta \sim \pi(\delta)$. This approach includes the task of optimizing under the point alternative $\delta = \delta_1$ as a special case when choosing a point prior $\pi(\delta) = \mathbf{1}_{\{\delta=\delta_1\}}$. The presented optimization under a weighted sum of effect sizes in Section 4 is a special case of the Bayesian approach with the prior $\pi(\delta) = p\mathbf{1}_{\{\delta=\delta_1\}} + (1-p)\mathbf{1}_{\{\delta=0\}}$. Optimizing not only under a single point effect size but under multiple effect sizes or even a continuous prior distribution helps to avoid overfitting on a single effect size. It can, therefore, also be seen as regularization of the design.

It is highly relevant to distinguish clearly between planning and reassessing two-stage designs. Flexible adaptations of the trial design can be performed as long as the conditional error principle holds. Note that determining the trial design during the planning stage by an optimization approach does not violate the conditional error principle. While the function $c_2(\cdot)$ serves as critical value for the stage-two test statistic, the conditional error function $A(\cdot)$ provides the critical value for the p -value of the second stage. Therefore, these two approaches are equivalent: When the optimal critical value function $c_2^*(\cdot)$ has been calculated, the corresponding conditional error function $A^*(\cdot)$ is given by

$$A^*(z_1) = 1 - \Phi(c_2^*(z_1)).$$

Hence, the conditional type I error rate can be computed for any value of z_1 . This allows changing all design elements as long as the conditional type I error rate is maintained. Therefore, it is still possible to modify the design during the ongoing trial and to react on new external information. However, leaving the path of the optimal design will inevitably decrease the efficiency with respect to the original objective criterion used during planning. We thus recommend to only deviate from the optimally planned design if this becomes necessary by new unforeseen information. Therefore, we agree with Bauer et al¹ who state that “[t]he question might arise if potential decisions made at interim stages might not be better placed to the upfront planning stage.” These considerations imply that much effort should be put in the planning of the trial design even if its adaptive character allows potential design modifications during the ongoing trial.

Finally, we would like to remark that the implementation of optimal designs in the software package *adoptr*³⁴ is a beneficial contribution to the community of adaptive design research. All examples in this manuscript have been implemented utilizing this package, and its generic character enables the users to solve their particular variant of the general optimization problem (1)-(3). We hope that the supply of such a software package encourages further research on optimal designs and in particular facilitates their application in clinical research.

ACKNOWLEDGMENTS

The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG) for supporting this research by grants KI 708/4-1 and RA 2347/4-1 and RUPRECHT KARLS UNIVERSITAET HEIDELBERG through Projekt DEAL and two reviewers for their comments that enormously helped to improve this work.

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 manuscript. All R code that was used to compute the examples on optimal designs is available as supplementary material.

ORCID

Maximilian Pilz  <https://orcid.org/0000-0002-9685-1613>

Kevin Kunzmann  <https://orcid.org/0000-0002-1140-7143>

Carolin Herrmann  <https://orcid.org/0000-0003-2384-7303>

Geraldine Rauch  <https://orcid.org/0000-0002-2451-1660>

Meinhard Kieser  <https://orcid.org/0000-0003-2402-4333>

REFERENCES

1. Bauer P, Bretz F, Dragalin V, König F, Wassmer G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Stat Med*. 2015;35(3):325-347. <https://doi.org/10.1002/sim.6472>.
2. Wassmer G, Brannath W. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer Series in Pharmaceutical Statistics. New York, NY: Springer International Publishing; 2016.
3. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994;50(4):1029-1041. <https://doi.org/10.2307/2533441>.
4. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999;55(4):1286-1290. <https://doi.org/10.1111/j.0006-341X.1999.01286.x>.
5. Cui L, Hung HMJ, Wang S-J. Modification of sample size in group sequential clinical trials. *Biometrics*. 1999;55(3):853-857. <https://doi.org/10.1111/j.0006-341X.1999.00853.x>.
6. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics*. 1995;51(4):1315-1324. <https://doi.org/10.2307/2533262>.
7. Vandemeulebroecke M. An investigation of two-stage tests. *Stat Sin*. 2006;16(3):933-951.
8. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977;64(2):191-199. <https://doi.org/10.1093/biomet/64.2.191>.
9. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35(3):549-556. <https://doi.org/10.2307/2530245>.
10. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman & Hall/ CRC Press; 1999.

11. Bauer P, König F. The reassessment of trial perspectives from interim data—a critical view. *Stat Med*. 2006;25(1):23-36. <https://doi.org/10.1002/sim.2180>.
12. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med*. 2011;30(28):3267-3284. <https://doi.org/10.1002/sim.4102>.
13. Jennison C, Turnbull BW. Adaptive sample size modification in clinical trials: start small then ask for more? *Stat Med*. 2015;34(29):3793-3810. <https://doi.org/10.1002/sim.6575>.
14. Bauer P. Adaptive designs: looking for a needle in the haystack—a new challenge in medical research. *Stat Med*. 2008;27(10):1565-1580. <https://doi.org/10.1002/sim.3090>.
15. Barber S, Jennison C. Optimal asymmetric one-sided group sequential tests. *Biometrika*. 2002;89:49-60. <https://doi.org/10.1093/biomet/89.1.49>.
16. Brannath W, Bauer P. Optimal conditional error functions for the control of conditional power. *Biometrics*. 2004;60(3):715-723. <https://doi.org/10.1111/j.0006-341X.2004.00221.x>.
17. Hsiao ST, Liu L, Mehta CR. Optimal promising zone designs. *Biom J*. 2019;61(5):1175-1186. <https://doi.org/10.1002/bimj.201700308>.
18. Schmitz N. *Optimal Sequentially Planned Decision Procedures*. New York, NY: Springer; 1993.
19. Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. *Biometrika*. 2006;93(1):1-21. <https://doi.org/10.1093/biomet/93.1.1>.
20. Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. *Stat Med*. 2006;25(6):917-932. <https://doi.org/10.1002/sim.2251>.
21. Lokhnygina Y, Tsiatis AA. Optimal two-stage group-sequential designs. *J Stat Plann Infer*. 2008;138(2):489-499. <https://doi.org/10.1016/j.jspi.2007.06.011>.
22. Pilz M, Kunzmann K, Herrmann C, Rauch G, Kieser M. A variational approach to optimal two-stage designs. *Stat Med*. 2019;38(21):4159-4171. <https://doi.org/10.1002/sim.8291>.
23. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989;10(1):1-10. [https://doi.org/10.1016/0197-2456\(89\)90015-9](https://doi.org/10.1016/0197-2456(89)90015-9).
24. Jung S-H, Lee T, Kim K, George SL. Admissible two-stage designs for phase II cancer clinical trials. *Stat Med*. 2004;23(4):561-569. <https://doi.org/10.1002/sim.1600>.
25. Mander AP, Thompson SG. Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. *Contemp Clin Trials*. 2010;31(4):572-578. <https://doi.org/10.1016/j.cct.2010.07.008>.
26. Shuster J. Optimal two-stage designs for single arm phase II cancer trials. *J Biopharm Stat*. 2002;12(1):39-51. <https://doi.org/10.1081/BIP-120005739>.
27. Wason JMS, Mander AP. Minimizing the maximum expected sample size in two-stage phase II clinical trials with continuous outcomes. *J Biopharm Stat*. 2012;22(4):836-852. <https://doi.org/10.1080/10543406.2010.528104>.
28. Wason JMS, Mander AP, Thompson SG. Optimal multistage designs for randomised clinical trials with continuous outcomes. *Stat Med*. 2012;31(4):301-312. <https://doi.org/10.1002/sim.4421>.
29. Banerjee A, Tsiatis AA. Adaptive two-stage designs in phase II clinical trials. *Stat Med*. 2006;25(19):3382-3395. <https://doi.org/10.1002/sim.2501>.
30. Englert S, Kieser M. Optimal adaptive two-stage designs for phase II cancer clinical trials. *Biom J*. 2013;55(6):955-968. <https://doi.org/10.1002/bimj.201200220>.
31. Kunzmann K., Kieser M. Optimal adaptive two-stage designs for single-arm trial with binary endpoint; 2016. arXiv:1605.00249.
32. Kunzmann K, Kieser M. Optimal adaptive single-arm phase II trials under quantified uncertainty. *J Biopharm Stat*. 2020;30(1):89-103. <https://doi.org/10.1080/10543406.2019.1609016>.
33. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020.
34. Kunzmann K., Pilz M. `adoptr`: adaptive optimal two-stage designs in R. R package version 0.4.1; 2020. <https://doi.org/10.5281/zenodo.4110773>.
35. Witzigmann H, Diener MK, Kienkötter S, et al. No need for routine drainage after pancreatic head resection: the dual-center, randomized, controlled PANDRA trial (ISRCTN04937707). *Ann Surg*. 2016;264(3):528-537. <https://doi.org/10.1097/SLA.0000000000001859>.
36. Heidelberg University Hospital Evaluation of robotic versus open partial pancreaticoduodenectomy – a randomised controlled trial (EUROPA); 2020. <https://www.drks.de/DRKS00020407>. Accessed January 21, 2021.
37. Slankamenac K, Graf R, Barkun J, Puhon MA, Clavien P-A. The comprehensive complication index: a novel continuous scale to measure surgical morbidity. *Ann Surg*. 2013;258(1):1-7. <https://doi.org/10.1097/SLA.0b013e318296c732>.
38. Slankamenac K, Nederlof N, Pessaux P, et al. The comprehensive complication index: a novel and more sensitive endpoint for assessing outcome and reducing sample size in randomized controlled trials. *Ann Surg*. 2014;260(5):757-762; discussion 762–763. <https://doi.org/10.1097/sla.0000000000000948>.
39. Statistical Principles for Clinical Trials The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) Statistical Principles for Clinical Trials - E9; 1998. <https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials>. Accessed January 21, 2021.
40. Wassmer G, Pahlke F. `rpact`: confirmatory adaptive clinical trial design and analysis R package version 3.0.3; 2020. <https://CRAN.R-project.org/package=rpact>

41. European Medicines Agency Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design; 2007. <https://www.ema.europa.eu/en/methodological-issues-confirmatory-clinical-trials-planned-adaptive-design>. Accessed January 21, 2021.
42. Herrmann C, Pilz M, Kieser M, Rauch G. A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation. *Stat Med*. 2020;39(15):2067-2100. <https://doi.org/10.1002/sim.8534>.
43. Brannath W, Bauer P, Posch M. On the efficiency of adaptive designs for flexible interim decisions in clinical trials. *J Stat Plann Infer*. 2006;136(6):1956-1961. <https://doi.org/10.1016/j.jspi.2005.08.014>.
44. Tsiatis AA, Mehta CR. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*. 2003;90(2):367-378. <https://doi.org/10.1093/biomet/90.2.367>.
45. Mehta CR, Patel NR. Adaptive, group sequential and decision theoretic approaches to sample size determination. *Stat Med*. 2006;25(19):3250-3269. <https://doi.org/10.1002/sim.2638>.
46. Posch M, Bauer P, Brannath W. Issues in designing flexible trials. *Stat Med*. 2003;22(6):953-969. <https://doi.org/10.1002/sim.1455>.
47. Lan KKD, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70(3):659-663. <https://doi.org/10.1093/biomet/70.3.659>.
48. Kim K, DeMets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*. 1987;74(1):149-154. <https://doi.org/10.1093/biomet/74.1.149>.
49. Kunzmann K., Pilz M., Herrmann C., Rauch G., Kieser M. The adoptr package: adaptive optimal designs for clinical trials in R; 2020. https://kkmann.github.io/adoptr/articles/adoptr_jss.html. Accessed January 21, 2021.
50. Fritsch FN, Carlson RE. Monotone piecewise cubic interpolation. *SIAM J Numer Anal*. 1980;17(2):238-246. <https://doi.org/10.1137/0717021>.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Pilz M, Kunzmann K, Herrmann C, Rauch G, Kieser M. Optimal planning of adaptive two-stage designs. *Statistics in Medicine*. 2021;40:3196–3213. <https://doi.org/10.1002/sim.8953>

APPENDIX

In this Appendix, it is briefly described how the general optimization problem to two-stage design (1)-(3) is solved in the R-package adoptr. For a detailed description of adoptr and its usage see its vignette.⁴⁹

Pilz et al²² solved a particular expression of this problem by applying the Euler Lagrange equations. In their work, the objective criterion was the expected sample size under the alternative, and constraints on maximum type I error rate and minimum power were imposed. This approach cannot be generalized to problems (1)-(3). A general solution would imply that the Euler Lagrange equations have to be defined from scratch for any expression of (1)-(3). This would be time-consuming and numerically unstable for complex expressions of this problem. Furthermore, any conditional constraint (eg, conditional power) lets the corresponding Lagrangian multiplier become a function. Then, no mathematical techniques as in Pilz et al²² can be applied to solve the Euler Lagrange equations.

Instead, a numerical approach is implemented in adoptr to solve (1)-(3). The continuation region $[c_f, c_e]$ is split into a set of k pivot points $c_f \leq z_1^{(1)} \leq \dots \leq z_1^{(k)} \leq c_e$. The stage-two functions $n_2(\cdot)$ and $c_2(\cdot)$ are evaluated only on these pivots and thus discretized. Consequently, a two-stage design is described by $2k + 3$ parameters $n_1, c_f, c_e, n_2(z_1^{(1)}), \dots, n_2(z_1^{(k)}), c_2(z_1^{(1)}), \dots, c_2(z_1^{(k)})$. All unconditional properties, as expected sample sizes or error rates, are evaluated by numerical integration of order k . Concretely, the method of choice for the numerical integration is Gauss Legendre quadrature. All conditional properties, as conditional power, are evaluated on the k pivot points. Cubic Hermite spline interpolation⁵⁰ of the stage two-functions $n_2(\cdot)$ and $c_2(\cdot)$ ensures the fulfillment of conditional constraints within the entire continuation region. This approach transforms the optimization problem (1)-(3) into a finite-dimensional one and can be solved by standard numerical libraries. Note that these libraries require the definition of lower and upper boundaries for the parameters to be optimized. Usually, there are internal definitions in adoptr as, for example, that the sample sizes must be at least 1. However, in complex optimization problems, a further restriction of the parameter space may allow faster convergence and helps to avoid convergence to strange solutions. In such cases, a manual definition of the lower and upper boundaries may be recommendable. Consequently, for some examples of this manuscript, specific lower and upper boundaries are included in the R code that is provided as supplemental material.