Aus dem Institut für Medizinische Biometrie und Informatik der
Universität Heidelberg
Abteilung Medizinische Biometrie
Geschäftsführender Direktor: Prof. Dr. sc. hum. Meinhard Kieser

Optimal Adaptive Designs for Early Phase II Trials in Clinical Oncology

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Kevin Kunzmann
aus Karlsruhe

2021

Dekan: Herr Prof. Dr. Hans-Georg Kräusslich
Doktormutter: Frau Prof. Dr. rer. nat. Annette Kopp-Schneider

'If one does not know to which port one is sailing, no wind is favourable.'

- Lucius Annaeus Seneca (ca. 4 BC to AD 65)

Contents

List of Figures ix

I. Introduction 1

1. Phase II Trials in Oncology 3
   1.1. Early clinical trials in oncology . . . 3
   1.2. The inadequacy of the standard binomial test . . . 5
   1.3. Simon's optimal two-stage designs . . . 7
      1.3.1. Example: binomial test and Simon's design . . . 8
   1.4. Adaptive interim analyses . . . 11
      1.4.1. Example: sample size recalculation . . . 13
   1.5. Aim and scope of this thesis . . . 16

II. Methods 17

2. Optimal Two-Stage Designs 19
   2.1. Notation . . . 19
      2.1.1. Notation example . . . 20
   2.2. Previous work . . . 22
   2.3. A novel solution via integer linear programming . . . 23
      2.3.1. Choice of maximal sample size and additional constraints . . . 25
   2.4. Other objective functions . . . 28

3. Optimisation Under Uncertainty 31
   3.1. Assessing performance under uncertainty . . . 31
   3.2. Prior choice . . . 33
   3.3. Power constraints under uncertainty . . . 36
   3.4. A utility-based approach . . . 39

4. Bayesian Inference 43
   4.1. Inference using the planning prior . . . 43
   4.2. Objective Bayesian inference . . . 44

5. Frequentist Inference 47
   5.1. The one-stage design situation . . . 47
   5.2. Frequentist inference in two-stage designs . . . 49
   5.3. Unbiased estimation . . . 50

5.4. Compatible frequentist inference 53

6. Unplanned Design Adaptations 57
6.1. Rationale for unplanned design adaptations 57
6.2. Notation for unplanned design adaptations 58
6.3. Optimal unplanned adaptations in stage two 59
6.4. Optimal unplanned adaptations in stage one 61
6.5. Previous work 63
6.6. Inference after unplanned design adaptations 63

7. Extension to the Continuous Case 65

III. Results 67

8. Examples: Optimal Two-Stage Designs 69
8.1. Generic optimal two-stage design 69
8.2. Alternative objective functions 72

9. Examples: Optimisation Under Uncertainty 75
9.1. Prior choice 75
9.2. Bayesian power constraints 78
9.3. Utility maximisation 82

10. Examples: Bayesian Inference 87
10.1. Posterior distribution and posterior mean estimator 87

11. Examples: Frequentist Inference 91
11.1. Point estimators, p values, and confidence intervals 91
11.2. Design incompatible p values 96
11.3. Compatible maximum likelihood estimator 98

12. Examples: Unplanned Adaptations 101
12.1. Unplanned adaptation in stage two 101
12.2. Unplanned adaptation in stage one 103

13. Examples: The Continuous Case 105

IV. Discussion and Summary 109

14. Discussion 111
14.1. Inadequacy of the single-stage binomial test 111
14.2. State of the art 111
14.3. An effective solution method 112
14.4. Optimal group-sequential designs 113
14.5. Optimisation under uncertainty 113

14.6. Optimisation *versus* unplanned recalculation 115
14.7. Inference 116
14.8. Conclusions 117

15. Summary 119

16. Zusammenfassung 121

17. Bibliography 123

18. Related Own Publications 129

A. Appendix 131
A.1. Software 131
A.2. Reproducibility of results 132
A.3. Acknowledgements 134
A.4. Affidavit - Eidesstattliche Versicherung 135

List of Figures

1.1. Power of the (randomised) one-stage binomial test. 6
1.2. Simon's design and fixed binomial test compared 10
1.3. Fixed binomial vs. adaptive recalculation vs. Simon's design 14
1.4. Conditional power comparison 15

8.1. Optimal design and regularity constraints 70
8.2. ILP complexity and solution time 71
8.3. Minimax design 73

9.1. Prior construction example 76
9.2. Prior effect on optimal expected sample size designs 77
9.3. Comparison of power constraints under uncertainty 79
9.4. Sensitivity of expected power with respect to the prior upper tail . . 80
9.5. Sensitivity of power to informative prior 81
9.6. Finding the matched utility design 83
9.7. Matched utility design 84
9.8. Sensitivity of utility-maximising design 85

10.1. Posterior mean inference 88
10.2. Jeffreys priors for two-stage designs 89

11.1. Frequentist and Bayesian point estimators 92
11.2. Estimator-induced p values 93
11.3. Frequentist and Bayesian interval estimators 95
11.4. Shan-design and estimator performance 97
11.5. Example for test incompatibility of p values 98
11.6. Compatible MLE for Shan-design 99

12.1. Unplanned adaptations - base design 102
12.2. Unplanned adaptations - partial re-calculated design 103

13.1. Continuous case example 106

Interactive Examples

All examples are also made available online as interactive notebooks, see `https://github.com/kkmann/optimal-binary-two-stage-designs` for details. More information on software and reproducibility can also be found in Appendix A.2. A subset of the methods discussed in this thesis is also available as interactive shiny app.

Part I.

Introduction

1. Phase II Trials in Oncology

1.1. Early clinical trials in oncology

The approval of a new drug is a difficult and lengthy endeavour. Clinical drug development programs in humans are usually divided in four phases (I to IV), each with its specific aims and scopes (U.S. Food and Drug Administration, 1997). In phase I, the focus lies on dose finding and obtaining a better understanding of the pharmacokinetics and the pharmacodynamics of a new compound. Phase II serves as an exploratory stage where different treatment regimes may be compared, dose-finding may be refined, and first evidence of therapeutic efficacy is established in comparably small clinical trials. Phase II trials play an important role in the drug development process. They serve as 'gatekeepers' intended to filter inactive substances before proceeding to phase III. This is crucial since the subsequent pivotal phase III trials are typically conducted as randomised controlled trials to establish definitive proof of efficacy. Consequently, phase III trials require substantially larger sample sizes and are thus more expensive and take longer to complete than any preceding studies. Finally, phase IV subsumes any post-approval long-term investigation of drug-use in practice and serves as a tool to collect sufficiently large amounts of safety data. This classification is in no way definitive and individual trials can be designed to fulfil a combination of objectives from different phases.

In early clinical oncology, phase II trials are often conducted as single-arm trials with the binary endpoint 'tumour response' (Ivanova *et al.*, 2016) as defined in the RECIST criterion (Therasse *et al.*, 2000). The primary reason for targeting the response rate rather than the gold standard of overall survival is trial duration and sample size: depending on the tumour type, it might be outright impossible to collect a sufficient number of events for a proper survival analysis in an early development phase. An alternative approach to this problem is the closer integration of a phase II trial within a larger phase III trial. One way of implementing such a tighter integration are seamless phase II/III trials that are conducted in two stages and allow an interim decision to proceed to stage two based on both observed tumour response rates and early survival data (Rufibach *et al.*, 2020). Yet, by far the most common approach in oncology is still the conduct of a dedicated phase II trial using tumour response as endpoint (Ivanova *et al.*, 2016). Here, the objective is to filter-out substances which fail to show sufficient tumour response before attempting to confirm efficacy with respect to overall survival in a subsequent larger phase III trial.

The rationale for using single-arm designs instead of the theoretically preferable randomised two- or multi-arm designs is at least two-fold. Firstly, a single-arm trial requires much smaller sample sizes which also leads to a speedier completion of the trial. The issue of small recruitment pools becomes more pressing as tumour dia-

gnostics progress and the question of efficacy has to be answered in ever smaller patient sub-populations. For instance, drugs targeting specific genetic variations or pathways might also be beneficial to patients with genetically similar tumours in indications which were not included in the initial approval of a drug. An example for such a drug is vemurafenib which orally inhibits BRAF (proto-oncogene B-Raf encoding gene) and works particularly well in tumours showing a specific mutation (BRAF V600 +) (Hyman *et al.*, 2015). Here, vemurafenib was initially approved for use in metastatic melanoma but its efficacy was subsequently investigated in different indications in a so-called 'basket trial' (Hyman *et al.*, 2015). This increase in potential applications of a single drug, which are often not foreseen when a first approval is sought, renders small but effective trial designs more important than ever. These designs may then either be implemented as stand-alone studies for extending approval of an existing drug or as part of a larger trial. Secondly, there are situations in which single-arm trials might be the only ethical option to proceed. For instance, when the biological mechanism of a new drug is such that a large improvement over the current gold standard can be expected or no such gold standard exists at all. In these situations, a randomised trial might be considered unethical due to the high *a priori* chance of treating a large group of individuals (the control arm) with an inferior compound or none at all. Still, there is evidence that the proportion of randomised trials in oncological phase II studies is increasing from about one third in 2005 to half in 2014 (Ivanova *et al.*, 2016). This trend might in part be due to the fact that phase II trials in oncology are increasingly used to obtain accelerated approval after phase II by the FDA which typically requires evidence obtained from at least one randomised trial:

> *'Using surrogate or intermediate clinical endpoints can save valuable time in the drug approval process. For example, instead of having to wait to learn if a drug actually extends survival for cancer patients, the FDA may approve a drug based on evidence that the drug shrinks tumours, because tumour shrinkage is considered reasonably likely to predict a real clinical benefit. In this example, an approval based upon tumour shrinkage can occur far sooner than waiting to learn whether patients actually lived longer. The drug company will still need to conduct studies to confirm that tumour shrinkage actually predicts that patients will live longer. [...] Where confirmatory trials verify clinical benefit, FDA will generally terminate the requirement. Approval of a drug may be withdrawn or the labelled indication of the drug changed if trials fail to verify clinical benefit or do not demonstrate sufficient clinical benefit to justify the risks associated with the drug (e.g., show a significantly smaller magnitude or duration of benefit than was anticipated based on the observed effect on the surrogate).'* (U.S. Food and Drug Administration, 2018).

In phase II oncological studies, typical per-arm patient numbers are relatively low with median sample size increasing only slowly from 39 (2005) to 45 (2014) (Ivanova *et al.*, 2016). This has important statistical consequences as typical asymptotic arguments are not reliable and the finite-sample properties of designs become more

important. Therefore, instead of invoking the central limit theorem for the observed response rate and assuming asymptotic normality, exact methods using the fact that the number of observed responses is binomially distributed are more adequate. In case of a one-stage design, this implies that testing the response rate of a new drug for superiority over a historical control reduces to an exact binomial test.

1.2. The inadequacy of the standard binomial test

Let X be the (random) number of responses out of n patients. Assume that all individuals share a common response probability, p. Then X is binomially distributed with parameters n and p. Let α be the maximal acceptable type one error rate for the null hypothesis $\mathcal{H}_0 : p \leq p_0$ where p_0 is often set to the typically well-known response rate under treatment as usual (TAU). This means that the objective of the trial is to demonstrate superiority in terms of the response rate under treatment over the currently established gold standard (TAU). The critical value c for the test decision to reject $\mathcal{H}_0$ in case $X > c$ is then chosen to be the smallest integer such that the probability to reject $\mathcal{H}_0$ is less than α under p_0, i.e. $\Pr_{p_0}[\, X > c\,] \leq \alpha$. Ideally, n is fixed at a large enough number to ensure a sufficient power of $1 - \beta$ at a point alternative $p_{\text{alt}} > p_0$, i.e. the sample size should be large enough to ensure $\Pr_{p_{\text{alt}}}[\, X > c\,] \geq 1 - \beta$. In early phase II trials, a relatively large maximal type one error rate is sometimes accepted (up to 10%) to allow sufficient power without increasing required sample size disproportionately (Simon, 1989). In the remainder of this thesis a more conservative combination of $\alpha = 5\%$ and $\beta = 80\%$ will be used as default if not otherwise indicated.

Two arguments render the standard binomial test unattractive for use in phase II trials. Firstly, the binomial test is ineffective whenever fixed error rates need to be satisfied. This is due to the discreteness of the underlying test statistic (the number of responses) (Simon, 1989). The discreteness implies that for any particular choice of p_0 and p_{alt}, the corresponding test with minimal sample size does not fully exhaust either of the acceptable error rates α and β. A standard statistical procedure for overcoming this inefficiency are 'randomised tests' where the final test decision is randomised such that the desired error rate constraints are met exactly (Lancaster, 1961). For any level-α binomial test, a randomised test can be constructed that fully exploits the permissible maximal type one error rate by randomising the test decision on the boundary such that the overall maximal type one error rate is exactly α. This means that the randomised test rejects the null when $X > c$ or $X = c$ and $U > c_{rand.}$, where U is an auxiliary random variable that follows a uniform distribution on $[\, 0, 1\,]$ and that is independent of the trial data. The quantity $c_{rand.}$ is an additional free parameter of the randomised test procedure. The random variable $\mathbf{1}_{U > c_{rand.}}$ determining the final outcome of the trial if $X = c$ can thus be thought of as a biased coin toss. The required randomisation probability to fit the maximal type one error rate constraint exactly and thus maximising power is

$$c^*_{rand.} := \frac{\alpha - \Pr_{p_0}[\, X > c\,]}{\Pr_{p_0}[\, X = c\,]}\,. \tag{1.1}$$

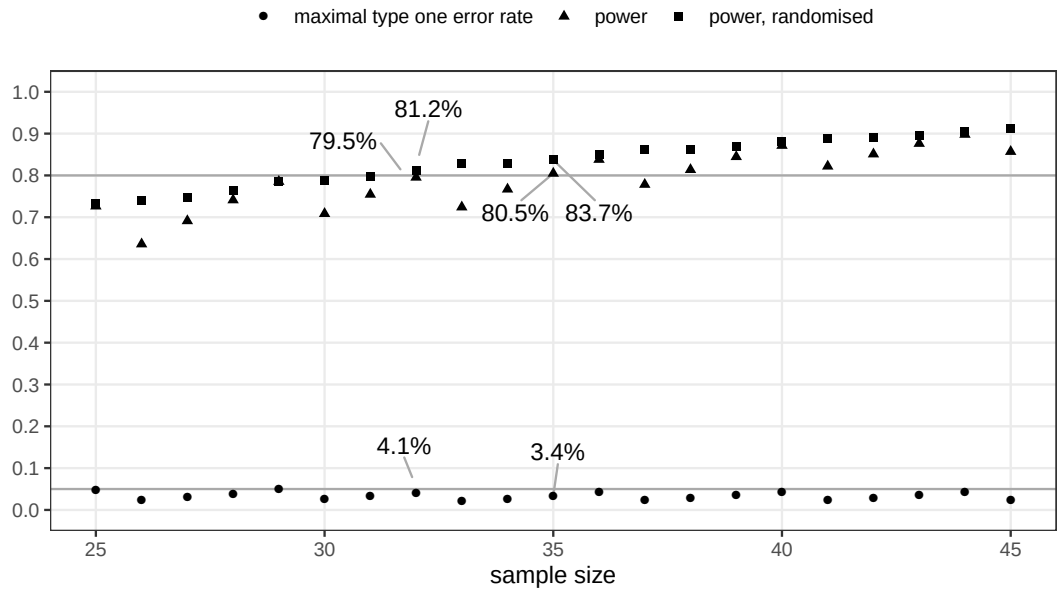

Figure 1.1.: Power at $p = 0.4$ and maximal type one error rate of binomial tests with $\mathcal{H}_0 : p \leq 0.2$ versus sample size of the tests; power of the corresponding optimal randomised test is also given (its maximal type one error rate is equal to $\alpha = 0.05$ by definition); sample size varies around the optimal value of 35 for a minimal power of 80%.

Here

$$\Pr_p[\, X = x \,] = \binom{n}{x} p^x \, (1-p)^{n-x} \tag{1.2}$$

is the binomial probability for x out of n responses with response probability p. A comparison between the randomised and non-randomised binomial test in terms of power and type one error rate is given in Figure 1.1 for $p_0 = 0.2$ and $p_{\text{alt}} = 0.4$. The binomial test with minimal overall sample size has substantially lower maximal type one error rate and slightly larger power than required. Lower-than-required error rates are, *per se*, no problem but the flip side is a needlessly large sample size. The randomised test exhausting the full permissible type one error rate for $n = 35$ achieves substantially larger power. Alternatively, the sample size can be reduced from 35 to 32 when using a randomised test. Although randomised tests nominally resolve the efficiency problem, the idea of resting the primary result of a multi-million-dollar trial on the equivalent of a biased coin toss is not acceptable in practice.

The second argument against a single-stage binomial test is that an upfront commitment to a relatively large sample size in a confirmatory setting is not always justified in phase II since there is typically large uncertainty about the assumed response probability p_{alt}. A design which allows early stopping for either futility (observed response rate is much lower than anticipated) or efficacy (observed response rate is much larger than anticipated) during the course of the trial is therefore an attractive option. To avoid error rate inflation due to optional stopping, such a design must consider the option of early stopping adequately when determining the bounds (Simon, 1989).

1.3. Simon's optimal two-stage designs

To overcome the ineffectiveness of the classical binomial test and its inability to reflect the large uncertainties about the anticipated response rate, phase II trials in oncology are often conducted using two-stage designs (Ivanova *et al.*, 2016). The aim of an interim analysis after $n_1 < n$ patients' responses have have been observed is to allow an early assessment of the response rate. In case of very few responses, the trial can be stopped early for futility and fewer patients are exposed to ineffective treatment within the study than in a classical binomial test. Should the observed response rate among the first n_1 patients be unexpectedly large, it is also possible to stop early for efficacy as long as the specified maximal type one error rate can be guaranteed. Two-stage designs are inherently more complex than the one-stage classical binomial test since they require the specification of more design parameters. Instead of just sample size and critical value, a two-stage design depends on the stage-one sample size n_1, potential early stopping boundaries, the stage-two sample size n_2, and the final rejection boundary after completing stage two. Even a two-stage design without early efficacy stopping has a minimum of four parameters and, consequentially, power and maximal type one error rate constraints alone are insufficient to determine all design parameters. For a one-stage design, the choice of objective criterion is unique since the test minimising the unique sample size n can unequivocally be considered optimal. The final sample size of a two-stage design, however, depends on the interim decision and the choice of objective criterion is thus no longer unique.

Simon's designs (Simon, 1989) are still among the most popular designs for single-arm binary two-stage trials in oncology (Ivanova *et al.*, 2016). Simon studied two-stage group-sequential designs for single-arm trials with binary endpoint minimising either the expected sample size on the boundary of the null hypothesis, p_0, or the maximal overall sample size, i.e., n_2. The idea of minimising the expected sample size under the null hypothesis is particularly attractive in oncological phase II trials to prevent the treatment of unnecessarily many individuals with if the new compound is indeed ineffective. Simon argued that

> *'[…t]he optimization criterion chosen here is not unique. One could minimize the expected sample size averaged with regard to a prior distribution for the true response probability p. Historically, however, most new regimens are not successful and, more importantly, optimizing the design for performance under the null hypothesis seems ethically appropriate.'* (Simon, 1989)

His idea of optimisation under a prior distribution will be revisited in Chapter 3. Simon also argued against allowing early stopping for efficacy:

> *'The ethical imperative for early termination occurs when the drug has low activity. When the drug has substantial activity […] there is often interest in studying additional patients in order to estimate the proportion, extent, and durability of response.'* (Simon, 1989)

This stance, however, is no longer shared unequivocally as speedy progression to the

next phase of a drug's clinical development program is often of great importance to stakeholders. For instance, Mander *et al.* (2010) stated that

> *'[...] [early] stopping for efficacy may save drug development time, bring useful treatments into clinical practice quicker, and should reduce costs.'*

Still, the classical Simon's design does not allow stopping for efficacy and the restriction to early stopping for futility renders the optimisation procedure particularly simple as any such design can be characterised by only four parameters. In Simon's original notation these are n_1, the sample size of the first stage, r_1, the boundary for rejecting the new drug after stage one (reject new drug, i.e. accept $\mathcal{H}_0$ if and only if $X_1 \leq r_1$), the final sample size if the trial does not stop early, n, and r, the final boundary for rejecting the new drug after completion of stage two. This means that Simon's design rejects $\mathcal{H}_0$ if and only if $X_1 + X_2 > r$ where X_i is the number of observed responses in stage i. This means that $r = c$ in the previous single-stage design notation. Note that Simon's notion of 'rejection' is in terms of the drug and therefore exactly opposite to the usual notion of rejection of the null hypothesis. For any given p_0, p_{alt}, α, and β Simon's optimal design (n_1, r_1, n, r) is the solution of

$$\underset{n_1, r_1, n, r}{\text{argmin}} : \quad n_1 \, \Pr_{p_0}[\, X_1 \leq r_1 \,] + n \, \Pr_{p_0}[\, X_1 > r_1 \,] \tag{1.3}$$

$$\text{subject to} : \quad \sum_{x_1 = r_1 + 1}^{n_1} \Pr_{p_0}[\, X_1 = x_1 \,] \, \Pr_{p_0}[\, X_2 > r - x_1 \,] \leq \alpha \tag{1.4}$$

$$\sum_{x_1 = r_1 + 1}^{n_1} \Pr_{p_{\text{alt}}}[\, X_1 = x_1 \,] \, \Pr_{p_{\text{alt}}}[\, X_2 > r - x_1 \,] \geq 1 - \beta \, . \tag{1.5}$$

Due to the discreteness of the problem and the small number of free parameters, the solution is easily obtained by an exhaustive brute force search over a grid defined by a minimal stage-one sample size and a maximal overall sample size. The maximal sample size for the grid search can be chosen as a multiple of the approximate sample size formula for the one-stage design

$$n \approx \left\lceil \, p_{\text{alt}} \, (1 - p_{\text{alt}}) \left(\frac{z_{1-\alpha} + z_{1-\beta}}{p_{\text{alt}} - p_0} \right)^2 \right\rceil . \tag{1.6}$$

Simon himself suggested a value of $1.5\, n$ (Simon, 1989).

1.3.1. Example: binomial test and Simon's design

To illustrate the characteristic differences between a one-stage binomial test and Simon's optimal design, consider the following situation. Assume that a new breakthrough therapy is to be assessed in a phase II trial and there is biological rationale to expect a substantial increase in the response rate over TAU. The response rate under TAU is well-established at $p_0 = 0.2$ and the anticipated response rate under treatment

is $p_{\text{alt}} = 0.4$. In this situation, a single-arm design is a suitable choice to establish initial evidence of drug activity in terms of an elevated response probability over TAU. Thus, a single-arm design with 80% power at a point alternative of $p_{\text{alt}} = 0.4$ and a maximal type one error rate of 5% is to be planned.

Both the one-stage binomial test and Simon's design for this situation (Simon, 1989) are compared in Figure 1.2. The custom plot in Figure 1.2 is specifically designed to compactly visualise all relevant information about a design and will be used throughout the reminder of this thesis to compare various designs. The first row of the panel depicts the two designs in the x_1/n plane. Each vertical bar gives the overall sample size depending on the observed number of stage-one responses x_1. The rejection region of the design corresponds to the black parts of the bars, the non-rejection region to the grey parts. The stage-one sample sizes are given behind the design names and are indicated as horizontal dotted lines directly in the plot. The numbers below each bar give the exact values of the stage-two critical value c_2 (top row) and the stage-two sample size n_2 (bottom row) for each x_1. Here, $n_2(x_1) := n(x_1) - n_1$ and $c_2(x_1) := c - x_1$. The configuration $c_2(x_1) = \infty$ and $n_2(x_1) = 0$ thus corresponds to early stopping for futility while $c_2(x_1) = -\infty$ and $n_2(x_1) = 0$ would correspond to early stopping for efficacy (not allowed in Simon's design). The second row of the panel shows plots for power and expected sample size of the designs as functions of all possible response probabilities in $[\,0, 1\,]$.

Note that a one-stage design can always be interpreted as a two-stage design with no continuation region. In the example situation, the binomial test enrols $n = 35$ subjects and rejects the null hypothesis whenever the observed number of responses is greater than $c = 11$. Simon's design requires a stage-one sample size of $n_1 = 13$. If $r_1 = 3$ or less responses are observed within the first n_1 subjects, the trial is stopped early for futility. Otherwise, the trial continues to a second stage until a total number of $n = 43$ subjects are enrolled. The null hypothesis is rejected after the final stage if the total number of responses in both stages is strictly greater than $c = r = 12$. The example shows that the omission of early efficacy stopping in Simon's designs leads to inefficiencies. In the unlikely case of 13 out of 13 responses in stage one, Simon's design still enrols 30 further patients for stage two although their response status has no effect on the final decision (reject $\mathcal{H}_0$).

The binomial test clearly suffers from the inefficiencies outlined in Section 1.2. It undershoots the target maximal type one error rate and exceeds the desired power. Simon's design, on the other hand, exploits the allowable error rates almost perfectly due to the larger number of free parameters. Another way of looking at this phenomenon is through the lens of randomised tests. In essence, Simon's design is a randomised test since its final decision depends on the (random) interim outcome. However, in contrast to the usual notion of a randomised test (Lancaster, 1961), the source of randomness is not external (a biased coin toss) but the trial-internal random outcome of stage one. Since the number of potential interim outcomes is small and the design's flexibility thus limited, the desired unconditional error levels cannot be matched perfectly but the advantage over an externally randomised test is that the final test decision is deterministic given the interim result $X_1 = x_1$.

In terms of sample size, the better exploitation of the permissible error rates and

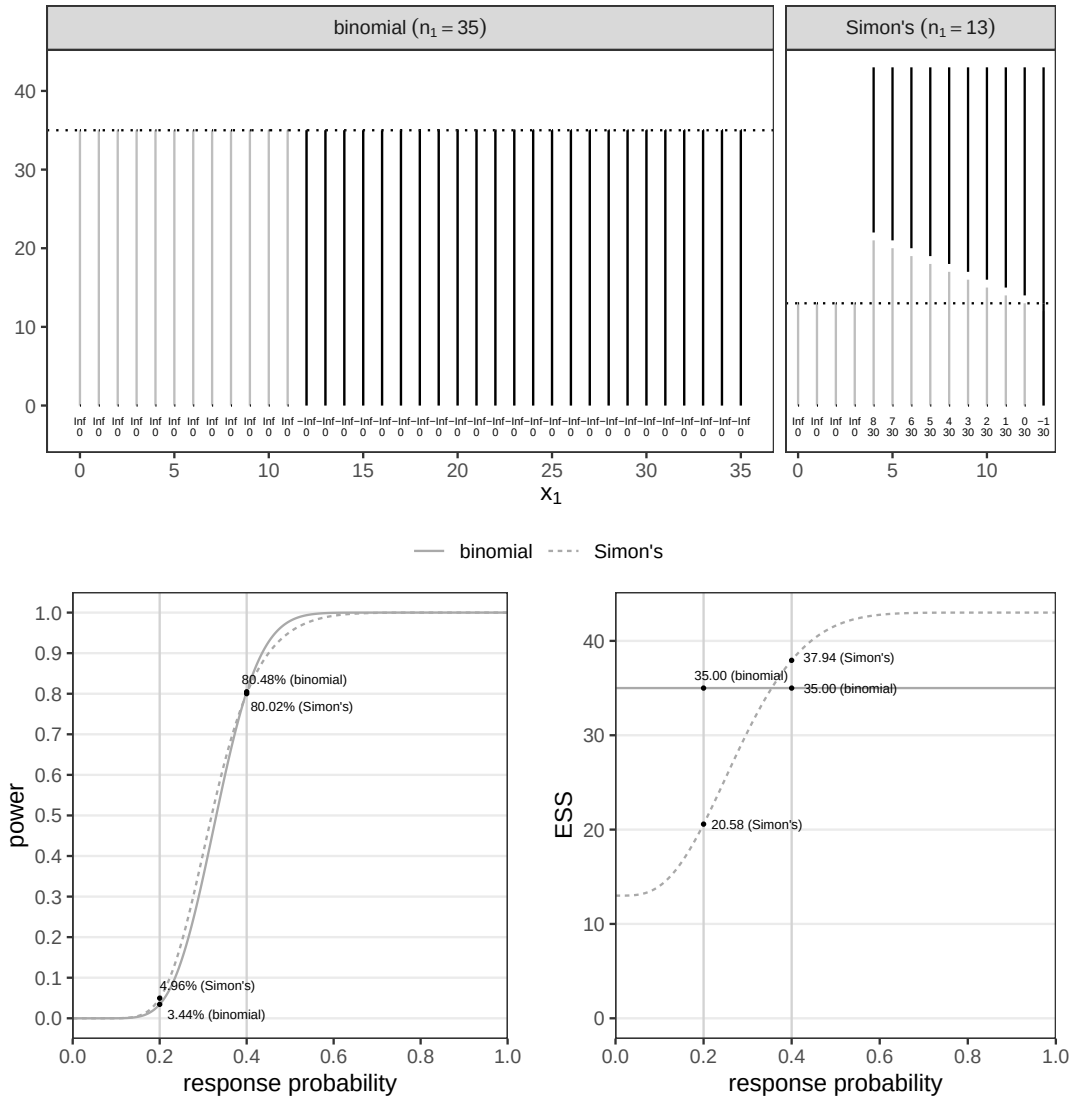

Figure 1.2.: Comparison of the classical binomial test and Simon's optimal design. The top panel shows the sample size and critical values for each design as functions of the observed number of interim responses x_1. The figuews under each vertical bar correspond to the stage-two critical values $c_2(x_1)$ (top row) and the stage-two sample sizes $n_2(x_1)$ (bottom row). The two bottom panels depict both designs' power and expected sample size (ESS) curves with exact values at the relevant points p_0 and p_{alt}. For a detailed explanation of the plot structure see Section 1.3.1.

the option to stop early for futility lead to a substantial reduction in expected sample size under p_0 for Simon's design. However, the expected sample size for true response rates close to and larger to p_{alt} is increased as compared to the binomial test. This is due to the maximal sample size of Simon's design being larger and the lack of an option for early efficacy-stopping. This comparison demonstrates that a two-stage design cannot be considered *per se* better than a one-stage test. Rather, its superiority is bound to the objective criterion of expected sample size on the boundary of the null hypothesis. An optimal two-stage design like Simon's greedily fits its free parameters to minimise the given objective. With respect to its specific objective, a two-stage design is always at least as good as the corresponding one-stage design because the feasible space of the optimisation problem is larger, i.e. because it has more free parameters. On the downside, this implies that the more flexible two-stage design 'overfits' the situation defined in the objective at the cost of performance under different assumptions on the true response rate.

1.4. Adaptive interim analyses

Although two-stage designs are able to react more flexibly to deviations from the original planning assumptions than the simple binomial test, their incorporation of interim results is fairly crude. Only a binary decision on early stopping is informed by the observed interim response rate. Instead, it could be desirable to adapt the entire second-stage, i.e. $n_2(x_1)$ and $c_2(x_1)$ during the interim analysis. In fact, the concept of two-stage designs was put forward by Simon (1989) for oncology in the same year that a more general methodology for adaptive interim analyses was proposed by Bauer (1989). Both approaches lead to designs where the final sample size depends on the observed interim outcome. However, their core principles are quite different. Simon considers his designs as the solution of an optimisation problem where the type one error rate restriction is simply incorporated as a constraint. The modification of the sample size is pre-specified before the start of recruitment. In contrast, Bauer transferred ideas from evidence synthesis to allow almost arbitrary *unplanned* adaptations of an ongoing trial without compromising strict type one error rate control. Over the years, the initial ideas of Bauer (1989) were substantially expanded beyond mere sample size adjustment (Bauer *et al.*, 2016) .

One way of formulating the approach to unplanned design adaptations dates back to ideas of R.A. Fisher who proposed to fuse two independent p values by means of a 'combination function' (Fisher, 1925). The idea can be mapped to the case of combining p values obtained from two consecutive stages of a trial as long as the second-stage p value, ρ_2, is conditionally independent of the one derived from the first stage, ρ_1. Instead of Fisher's originally proposed combination function

$$\mathrm{f}_{\text{Fisher}}(\rho_1, \rho_2) = \rho_1 \cdot \rho_2 \tag{1.7}$$

(Fisher, 1925; Bauer *et al.*, 1994), many clinical trials use the inverse normal combin-

ation function

$$\mathrm{f}_{\text{IN}}(\rho_1, \rho_2) = 1 - \Phi\left(\sqrt{\frac{n_1}{n_1+n_2}}\,\Phi^{-1}(1-\rho_1) + \sqrt{\frac{n_2}{n_1+n_2}}\,\Phi^{-1}(1-\rho_2)\right) \tag{1.8}$$

proposed by Lehmacher *et al.* (1999). Defining a futility boundary b_f and an efficacy boundary $b_e < b_f$ for ρ_1 and an overall critical value b for the combined p values, the type one error rate under the null hypothesis is given by

$$\Pr_{p_0}[\, P_1 < b_e \,] + \Pr_{p_0}\left[\, \mathrm{f}_{\text{IN}}\left(P_1, P_2\right) \leq b\,,\; P_1 \in [\, b_e, b_f \,]\,\right] \tag{1.9}$$

where P_i is the (random) stage-i p value. Here, the trial stops early for futility if $P_1 > b_f$ and for efficacy if $P_1 < b_e$. As long as the conditional distribution of $P_2 \,|\, P_1$ under the null is stochastically at least as large as a uniform distribution on $[\,0,1\,]$, the type one error rate is bounded from above by

$$b_e + \int_{b_e}^{b_f} \mathbf{1}_{\mathrm{f}_{\text{IN}}(\rho_1,\rho_2)\leq b}\; \mathrm{d}\rho_1\; \mathrm{d}\rho_2\;. \tag{1.10}$$

Quantities such as the stage-two sample size can then be adapted freely as long as the decision boundaries in the space of stage-wise p values and the pre-specified combination function are maintained (Bauer *et al.*, 1999).

Alternatively, a 'conditional error function' can be used to express invariant rejection boundaries in the stage-wise p value space (Proschan *et al.*, 1995). Here, a conditional error function $\mathrm{CE} : [\,0,1\,] \mapsto [\,0,1\,]$ is a function satisfying the defining constraint

$$\int_0^1 \mathrm{CE}(\rho_1)\,\mathrm{d}\rho_1 \leq \alpha\;. \tag{1.11}$$

Upon observing $P_1 = \rho_1$, the second stage p value ρ_2 can then be tested against the conditional error level $\mathrm{CE}(\rho_1)$ and by similar arguments as above the procedure again maintains type one error rate control irrespective of unplanned interim modification to the trial. In fact, both approaches are equivalent as they merely define a rejection region in the (ρ_1, ρ_2)-space and rely on the conditional independence as well as uniformity under p_0 of the stage-wise p values (Vandemeulebroecke, 2006). Any initial design $\mathcal{D}$, implicitly defines a conditional error function

$$\mathrm{CE}_{\mathcal{D}}(x_1) := \Pr_{p_0}\left[\,\text{reject } \mathcal{H}_0 \,|\, X_1 = x_1, \mathcal{D}\,\right]\;. \tag{1.12}$$

The definition is given in terms of the observed test statistic x_1 but can easily be mapped to $[\,0,1\,]$ by computing the stage-one p values. Also, for discrete test statistics, this implicit conditional error function is only partially defined since only a finite set p values in $[\,0,1\,]$ can be observed.

To conduct an unplanned recalculation after observing $X_1 = x_1$ while maintaining strict type one error rate control, a new design $\mathcal{D}'$ simply needs to fulfil

$$\mathrm{CE}_{\mathcal{D}'}(x_1) \leq \mathrm{CE}_{\mathcal{D}}(x_1)\;. \tag{1.13}$$

A further generalisation of the conditional error function approach was put forward by Müller and Schäfer (Müller *et al.*, 2004). Their 'conditional error principle' states that, for overall type one error control, it is sufficient if the conditional error of a modified design is smaller or equal to the conditional error of the original design given the data observed so far. This principle is slightly more general than the conditional error function approach in that the conditional error of the old design varies depending on the time point of the unplanned interim analysis. A formal probabilistic proof that this principle indeed holds under mild technical assumptions was provided by Brannath *et al.* (2012). Due to its generality, the conditional error principle will be used throughout this thesis whenever unplanned interim analyses are considered.

1.4.1. Example: sample size recalculation

A comparison of Simon's optimal design and an adaptive recalculation design with the one-stage binomial test for the example situation discussed in Section 1.3.1 illustrates the key differences. For ease of comparison, the unplanned interim analysis is conducted at the same time as in Simon's design, i.e. when the outcome of 13 individuals are observed.

A popular method for sample size adaptation is based on conditional power at the originally assumed p_{alt} (Proschan *et al.*, 1995). During the interim analysis, the conditional power of the design given the data collected before the interim analysis is computed. Just as unconditional power, conditional power for any particular $X_1 = x_1$ is a function of the response probability p

$$\text{CP}(x_1, p) := \Pr_p \left[X_2 > c_2(x_1) \,|\, X_1 = x_1 \right] . \tag{1.14}$$

For any x_1, the conditional error is simply another point on the conditional power curve evaluated at p_0, i.e., $\text{CE}(x_1) = \text{CP}(x_1, p_0)$. Whenever conditional power drops below 80% during the interim analysis, the sample size is increased until the target power is met again. The rationale for this procedure is that even under $p = p_{\text{alt}}$ a trial might, by chance, start with disproportionately many non-responders, i.e. ${}^{x_1}/_{n_1} < p_{\text{alt}}$. As a consequence, the conditional power for the second stage could drop below the originally planned 80% without adjusting the sample size accordingly. Invoking the conditional error principle for the single-stage binomial test in this particular situation ($n = 35, c = 11$) after the same number of responses as in Simon's design ($n_1 = 13$) implies that the modified stage-two sample size and critical bounds $n_2'(x_1)$ and $c_2'(x_1)$ for each x_1 are obtained as solution of

$$\underset{n_2'(x_1),\, c_2'(x_1)}{\text{argmin}} : \quad n_2' \tag{1.15}$$

$$\text{subject to :} \quad \Pr_{p_0}[\, X_2' > c_2' \,] \leq \Pr_{p_0}[\, X_2 > 11 - x_1 \,] \tag{1.16}$$

$$\Pr_{p_{\text{alt}}}[\, X_2' > c_2' \,] \geq 0.8 \tag{1.17}$$

$$22 \leq n_2' \leq 70 - 13 = 57, \tag{1.18}$$

where $X_2 \sim \text{Binomial}(\, 25,\, p)$ and $X_2' \sim \text{Binomial}(\, n_2'(x_1),\, p)$. The maximal sample size limit is arbitrarily set to twice the sample size of the original single-stage test (70)

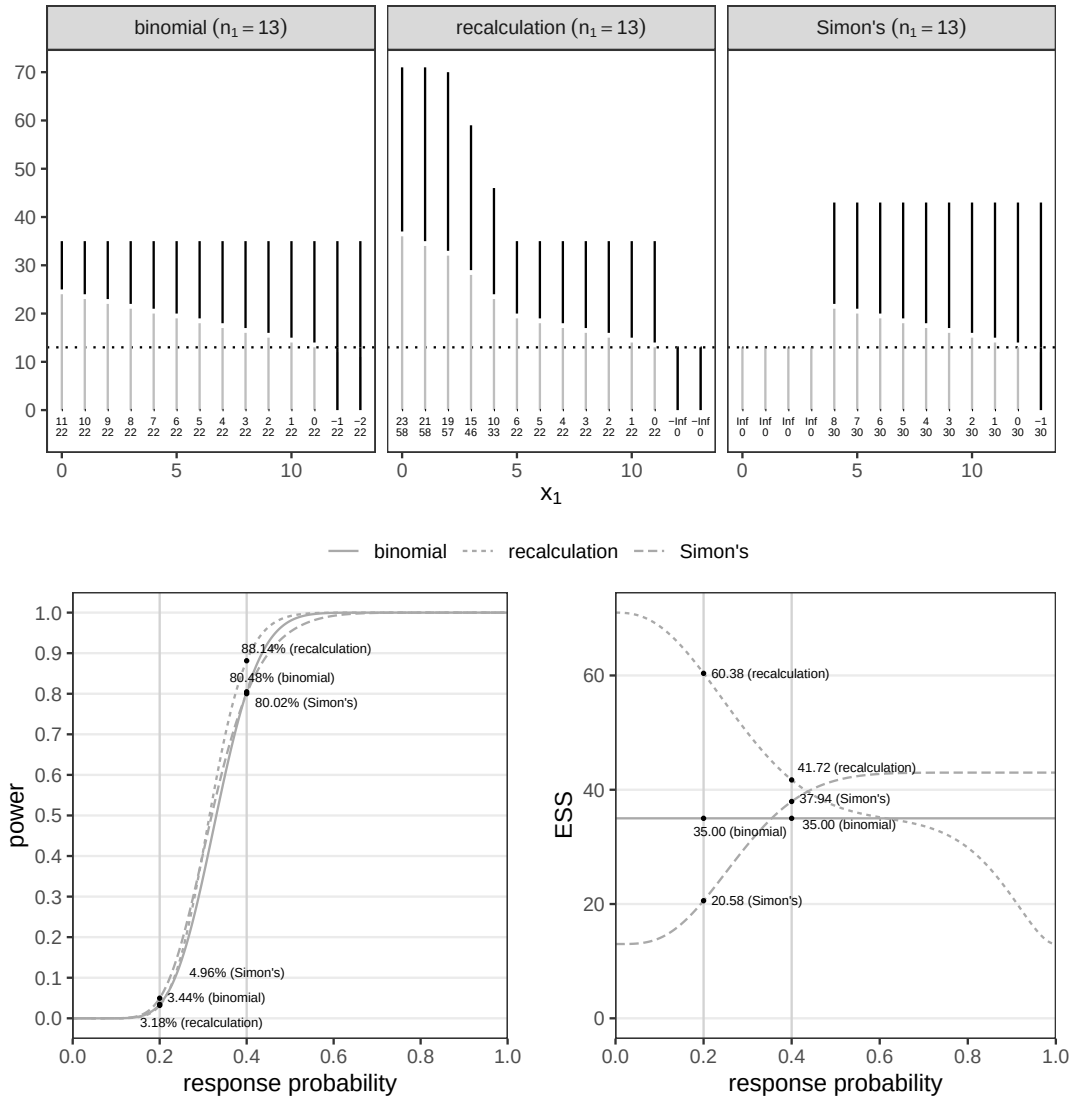

Figure 1.3.: Comparison of adaptive recalculation based on conditional power of the standard binomial test after 13 individuals with the original test and Simon's design (see also example in Section 1.3.1).

to prevent impractically large increases for very low response rates. The lower limit of 22 for $n_2(x_1)$ encodes the fact that that the sample size should only be increased. During the recalculation, implied early stopping for futility can be incorporated to reduce the sample size for $x_1 = 12$ and $x_1 = 13$.

Even though the described adaptation procedure is unplanned, the design resulting from its binding application for all x_1 can be studied. The design is obtained by pre-calculating all potential adaptations and is compared with the original binomial test and Simon's design in Figure 1.3. To facilitate comparison, the one-stage binomial test is re-interpreted as a two-stage design with $n_1 = 13$.

Clearly, the recalculated binomial test shows very different characteristics than Simon's design. The sample size is *increased* for small observed response rates whereas Simon's design favours aggressive early stopping for futility to *reduce* the expected sample size under the null hypothesis. Consequentially, the expected sample size

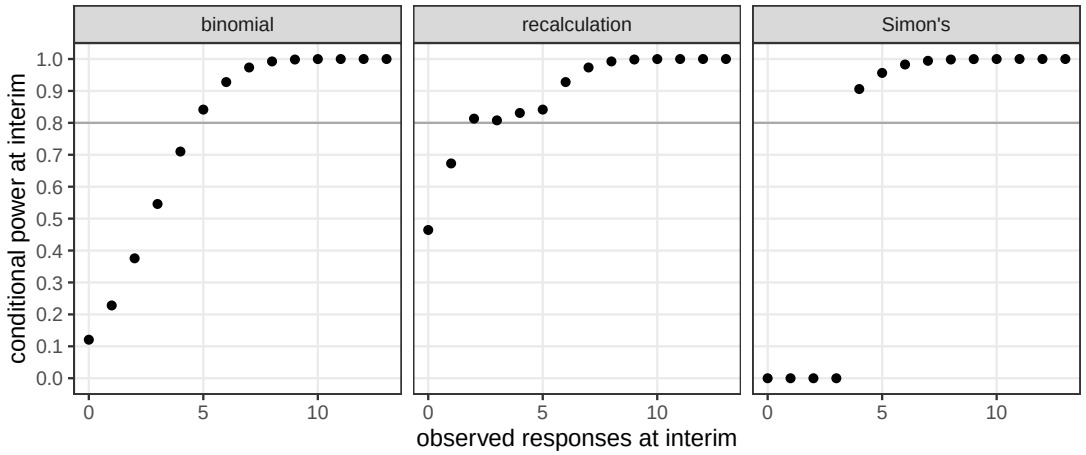

Figure 1.4.: Conditional power of the unmodified binomial test, the recalculated binomial test, and Simon's design at the interim analysis after $n_1 = 13$ individuals.

profile is almost the exact opposite of Simon's design with large increases over the fixed binomial test for low response rates and a reduction towards very high response rates due to early efficacy stopping. Since the conditional power is only ever increased under the recalculation, the overall power is also increased substantially over the targeted 80%. Due to the application of the conditional error principle, the maximal unconditional type one error rate of the recalculated design is even lower than the one of the one-stage binomial test.

The effect of the recalculation on conditional power is shown in Figure 1.4. The objective of stabilising the chances of a successful rejection in stage two is met for the recalculated design unless the sample size limit is hit ($x_1 = 0, 1$). Simon's design avoids situations with low conditional power since the objective criterion favours early stopping for futility in these cases.

This example demonstrates how different objectives in the planning of a clinical trial can lead to quite different characteristics. The discrepancy is mostly driven by the fact that Simon's objective minimises sample size exactly for those interim results where the recalculation heuristic based on conditional power needs to increase sample size. From the perspective of an unplanned adaptation it makes sense that the recalculated design considers each interim result $X_1 = x_1$ completely independently of other possible interim outcomes. However, since it is optimised unconditionally, Simon's design eliminates the need to recalculate based on conditional power by more aggressive early stopping while still maintaining the desired overall power and naturally avoiding situations in which the conditional power drops too low during the interim analysis.

1.5. Aim and scope of this thesis

The aim of this thesis is to develop a rigorous and consistent framework for all statistical aspects of planning and evaluating a single-arm phase II trial with binary endpoint 'tumour response to treatment'. This includes guidance on the definition of a situation-specific objective criterion under planning uncertainty, methods to react flexibly to new trial-external evidence that might arise during the course of the trial, and inference after concluding the trial.

To this end, a novel numerical approach is presented which makes the global optimisation of such design feasible in practice and improves existing approaches in terms of both flexibility and speed. The problem of incorporating *a priori* uncertainty about the true effect size in the planning process is discussed in detail taking a Bayesian perspective on quantifying uncertainty about the unknown response probability p is taken. Subsequently, the close interplay between point estimation, p values, confidence intervals, and the final test decision is illustrated and a framework is developed which allows consistent and efficient inference in binary single-arm two-stage designs. Finally, issues are addressed that may arise during the implementation of the proposed methods in practice. In particular, the problem of unplanned design modifications is revisited and the distinction between pre-specified adaptations within optimal two-stage designs and unplanned adaptations of ongoing designs discussed in more depth.

Part II.

Methods

2. Optimal Two-Stage Designs

In recent years, the approach of Simon to optimise group-sequential two-stage designs for binary endpoints with respect to a suitable criterion was extended by several authors to more flexible classes of designs (Banerjee *et al.*, 2006; Englert *et al.*, 2013; Shan *et al.*, 2016; Kunzmann *et al.*, 2016). This chapter discusses the practical difficulties that arise from this class of optimisation problems. A novel solution method is presented which improves over existing approaches both in terms of speed and flexibility.

2.1. Notation

In the following, notation encompassing all binary single-arm one- and two-stage designs is introduced. Some of the definitions were already discusses in the preceding chapter but, in the interest of clarity, all relevant terms are introduced again in context. The notation has evolved over a series of publications related to the topic and is loosely based on (Kunzmann *et al.*, 2016, 2017a,b, 2020a).

Let n_1 be the stage-one sample size of a two-stage design and $X_1 \sim \text{Binomial}(\, n_1,\, p)$ a random variable representing the number of observed responses (successes) in stage one under a new treatment of interest. Here, $p \in [\, 0, 1\,]$ denotes the unknown true response probability under the new treatment. The simple binomial model assumes that the response probability is constant within the population of interest. This assumption is only plausible within a sufficiently narrow target population. For larger target populations, regression models might be more appropriate to take heterogeneity of treatment effect via subject-specific covariates into account. For many situations in early oncological phase II trials, the assumptions of a shared response probability p is plausible since the biological mechanism through which a new compound affects tumour response is often well-understood. The target population can then be expected to be fairly homogeneous in terms of their response to treatment. In fact, one of the reasons that makes small, effective trials so important in this setting is the explosion of diagnostic information leading to ever smaller target populations (see Section 1.1).

Let $1 \leq n_{\text{max}} < \infty$ be the maximal feasible sample size for the trial. In practice, n_{max} will be much lower than 150 for early phase II trials in oncology since treatment and thus trials are expensive. Only in a subsequent phase III study, larger sample sizes are realised (Ivanova *et al.*, 2016). Let

$$n_2 : \{0, \ldots, n_1\} \rightarrow \{0, 1, \ldots, n_{\text{max}} - n_1\} \tag{2.1}$$

be a discrete function mapping from the observed number of stage-one responses $X_1 = x_1$ to the stage two sample size $n_2(x_1)$. Here, $n_2(x_1) = 0$ corresponds to the case of stopping the trial after the first stage. Let X_2 denote the stage-two number of

responses. The random variable X_2 also follows a binomial distribution conditional on continuation to stage two, i.e., $X_2 \,|\, X_1 = x_1, n_2(x_1) > 0 \sim \text{Binomial}(\, n_2(x_1),\, p)$. Conditional on $n_2(X_1) = 0$, the distribution reduces to a point mass on 0. The final test decision is governed by another discrete function

$$c_2 : \{0, \ldots, n_1\} \to \{-\infty, 0, 1, \ldots, n_{\text{max}} - n_1 - 1, \infty\} \tag{2.2}$$

mapping x_1 to the stage-two critical value $c_2(x_1)$. After observing the response status of $n(X_1) := n_2(X_1) + n_1$ individuals, the null hypothesis $\mathcal{H}_0 : p \leq p_0$ is rejected if and only if $X_2 > c_2(X_1)$. In the case of early stopping, c_2 is ill-defined but can be made uniquely identifiable by restricting its values to $\{-\infty, \infty\}$ and requiring that

$$|c_2(x_1)| = \infty \Leftrightarrow n_2(x_1) = 0\,. \tag{2.3}$$

It follows that $c_2(x_1) = \infty$ corresponds to early stopping for futility and similarly that $c_2(x_1) = -\infty$ encodes early stopping for efficacy. For the sake of completeness, $c(x_1) := c_2(x_1) + x_1$ can be defined. Let also $X := X_1 + X_2$. An equivalent definition of the rejection region in terms of the overall number of responses is then given by $X > c(X_1)$.

Any single-arm two-stage binary design for testing $\mathcal{H}_0$ can thus be seen as a three-tuple

$$\mathcal{D} := \big(\, n_1, n_2(\cdot), c_2(\cdot)\,\big) \tag{2.4}$$

where n_1 can be omitted since it is implicit in the domain of definition of both $n(\cdot)$ and $c(\cdot)$.

Within this framework, group-sequential designs are a special case with constant stage-two sample size upon continuation. One-stage designs are also encompassed although there is no unique way of representing them. The conceptually most straightforward way is to formalise them as two-stage designs with $n_2(x_1) = 0$ (see Figure 1.2) but the same design can also be seen as a two-stage design (see Figure 1.3).

Early stopping is entirely optional and can be disallowed by restricting the codomain of c_2 to exclude either $-\infty$ (no early stopping for efficacy), ∞ (no early stopping for futility), or both.

2.1.1. Notation example

To make things more tangible, consider again Simon's optimal design for testing $\mathcal{H}_0 : p \leq 0.2$ against the point alternative $p_{\text{alt}} = 0.4$ with maximal type one error rate of 5% and a power of 80% discussed in Section 1.3.1. The design has a stage-one sample size of 13, a maximal sample size of 43, rejects the null hypothesis if more than 12 responses are observed in stage one, and stops early for futility when 3 or less responses are observed in stage one (Simon, 1989). This design can be represented in the notation introduced above as $\mathcal{D}_{\text{Simon}} := \big(\, 13, n_2(\cdot), c_2(\cdot)\,\big)$ with the stage-two sample size function $n_2(x_1) = 30 \cdot \mathbf{1}_{x_1 > 3}(x_1)$ and the stage-two critical value function $c_2(x_1) = 12 - x_1 + \infty \cdot \mathbf{1}_{x_1 \leq 3}(x_1)$. Here the convention $0 \cdot \infty := 0$ is adopted.

Simon's criterion is expected sample size on the boundary of the null hypothesis. Early stopping for efficacy is not allowed in Simon's original optimal design and the

solution is restricted to group-sequential designs, i.e., the stage-two sample size on the continuation region must be constant. The design $\mathcal{D}_{\text{Simon}}$ is thus the solution of the following discrete optimisation problem

$$\underset{n_1,\, n_2(\cdot),\, c_2(\cdot)}{\text{argmin}} : \quad \sum_{x_1=0}^{n_1} \Pr_{p_0}[\, X_1 = x_1 \,]\, n(x_1) \tag{2.5}$$

$$\text{subject to :} \quad \Pr_{p_0}[\, X_2 > c_2(X_1) \,] \leq \alpha \tag{2.6}$$

$$\Pr_{p_{\text{alt}}}[\, X_2 > c_2(X_1) \,] \geq 1 - \beta \tag{2.7}$$

$$n_2(x_1) = 0 \Leftrightarrow |c_2(x_1)| = \infty$$

$$\forall\, x_1 = 0, \ldots, n_1 \tag{2.8}$$

$$c_2(x_1) = \infty \Rightarrow c_2(x_1 - 1) = \infty$$

$$\forall\, x_1 = 1, \ldots, n_1 \tag{2.9}$$

$$c_2(x_1) \neq -\infty$$

$$\forall\, x_1 = 0, \ldots, n_1 \tag{2.10}$$

$$n_2(x_1) > 0 \Rightarrow c_2(x_1 + 1) = c_2(x_1) + 1$$

$$\forall\, x_1 = 0, \ldots, n_1 - 1 \tag{2.11}$$

$$n_2(x_1) > 0 \Rightarrow n(x_1 + 1) = n(x_1)$$

$$\forall\, x_1 = 0, \ldots, n_1 - 1. \tag{2.12}$$

Here, constraints (2.6) and (2.7) ensure the type one and type two error rate constraints[1]. Constraint (2.8) enforces consistency between $n_2(\cdot)$ and $c_2(\cdot)$ in terms of early stopping. Constraint (2.9) ensures that the stopping-for-futility region is contiguous and, if present, starts a $x_1 = 0$. Constraint (2.10) excludes the possibility of early stopping for efficacy. Constraint (2.12) forces $n_2(\cdot)$ to be locally constant on the continuation region restricting the feasible space to group-sequential designs. Similarly, constraint (2.11) makes sure that $c(\cdot)$ is locally constant, an additional constraint used by Simon to make the search space small enough for an exhaustive grid search.

The above given formulation of Simon's optimisation problem might be seen as needlessly complex since any Simon's design can alternatively be described by merely four numbers (n_1, r_1, r, and n, see Section 1.3). Instead, the proposed notation increases the complexity to an almost arbitrary number of parameters by considering the discrete functions $n_2(\cdot)$ and $c_2(\cdot)$. Generic discrete functions are fully characterised by the function values on their (here) finite domain. The number of overall parameters is thus only limited by the fact that $n_1 \leq n_{\text{max}}$. For, e.g., $n_{\text{max}} = 100$, this is

[1] There is no formal proof that the constraint on the boundary of the null hypothesis is sufficient to ensure strict type one error rate control but this can be easily checked for any solution by inspecting the power function for monotonicity.

implies $2 \cdot (n_{\text{max}} + 1) + 1 = 201$ parameters - a roughly 50-fold increase over the four parameters in a classical Simon's design. This increased flexibility is then controlled via a set of constraints in the above stated optimisation problem. For instance, the omission of constraint (2.9) would allow designs with non-contiguous stopping for futility regions - due to the discreteness of the underlying binomial distribution such implausible solutions can not be ruled out entirely (Kunzmann *et al.*, 2016). However, the generic formulation also makes transparent that Simon's optimisation problem is needlessly restrictive. Constraint (2.11) is merely required to reduce $c_2(\cdot)$ to a single global critical value for the overall number of responses and could be dropped even if a group-sequential solution was required. This and constraint (2.12) make the connection to optimal *generic* two-stage designs with response-adaptive $n_2(\cdot)$ and $c_2(\cdot)$ on the continuation region evident. Dropping these two constraints from the problem formulation will result in an optimal generic two-stage design where the stage-two sample size and critical value functions are allowed to adapt to the individual interim outcomes $X_1 = x_1$ instead of the mere decision to continue to the second stage or not. Fewer constraints imply a larger feasible space for solutions and thus the performance (expected sample size under p_0) of the resulting design must be at least as good as the corresponding group-sequential one. However, dropping these two constraints also expands the feasible space through combinatorial explosion beyond what could realistically be explored via a brute-force search and thus requires more sophisticated solution methods.

2.2. Previous work

The core issues with problems of the class (2.5)-(2.12) are the discreteness of the underlying parameters, the complexity of the additional constraints (consistent early stopping etc.), and the fact that the dimensionality of the discrete functions $n_2(\cdot)$ and $c_2(\cdot)$ depends on n_1 and thus also the number of parameters.

Up to now, authors approached these issues by obtaining solutions conditional on n_1 and then implementing a grid search over a plausible (small) range of n_1 values. Banerjee *et al.* (2006) overcame the discreteness issue by solving a relaxation where $n_2(\cdot)$ and $c_2(\cdot)$ where allowed to take real values. The resulting smooth problem was addressed via backward induction before discretising the real valued-solutions by rounding to the nearest integer. Consequently, their solutions are only approximate and it is not clear how additional discrete constraints would be handled within this framework. Since sample sizes tend to be small in early phase II trials, the relaxed solutions may be ineffective or even violate the required error rate constraints.

Englert *et al.* (2013) implemented an exhaustive search of the space of designs via a custom Branch & Bound algorithm (Nemhauser *et al.*, 1988). Their solution strategy, however, is still limited to relatively small designs. They needed to impose an additional technical constraint which, together with relatively strict limits on n_{max} (+10% as compared to the corresponding maximal sample size of Simon's optimal design), reduced the size of the solution space and makes a naïve Branch & Bound search feasible. In particular, they restricted the search space to designs for which the conditional type one error rate on the boundary of the null hypothesis is monotonically

increasing, i.e.,

$$\mathrm{CE}[\,x_1 \,|\, \mathcal{D}\,] \geq \mathrm{CE}[\,x_1 - 1 \,|\, \mathcal{D}\,] \quad \forall x_1 = 1, 2, \ldots, n_1. \tag{2.13}$$

Although this constraint may be intuitively sensible, it is by no means a necessary one.

2.3. A novel solution via integer linear programming

Consider the following slightly more general version of the problem leading to Simon's optimal (group-sequential) design

$$\underset{n_1,\, n_2(\cdot),\, c_2(\cdot)}{\operatorname{argmin}} : \quad \sum_{x_1=0}^{n_1} \Pr_{p_0}[\,X_1 = x_1\,]\, n(x_1) \tag{2.14}$$

$$\text{subject to :} \quad \Pr_{p_0}[\,X_2 > c_2(X_1)\,] \leq \alpha \tag{2.15}$$

$$\Pr_{p_{\text{alt}}}[\,X_2 > c_2(X_1)\,] \geq 1 - \beta \tag{2.16}$$

$$n_2(x_1) = 0 \Leftrightarrow |c_2(x_1)| = \infty, \forall\, x_1 = 0, \ldots, n_1 \tag{2.17}$$

$$c_2(x_1) = \infty \Rightarrow c_2(x_1 - 1) = \infty, \forall\, x_1 = 1, \ldots, n_1 \tag{2.18}$$

$$c_2(x_1) = -\infty \Rightarrow c_2(x_1 + 1) = \infty, \forall\, x_1 = 0, \ldots, n_1 - 1\,. \tag{2.19}$$

Here, merely the restriction to a group-sequential design is dropped and early stopping for efficacy is allowed. The only non-essential constraints are those for contiguous early stopping. Kunzmann *et al.* (2016) studied the effects of dropping these 'nicety constraints' in detail. Due to the discreteness of the problem, in some rare cases a configuration with non-contiguous early stopping may actually be slightly more effective but the performance differences are negligible. In practice, a design with non-contiguous early stopping rules would hardly be accepted. The following discussion is thus restricted to designs *with* contiguous stopping.

For any given n_1 the objective function is a linear function in the stage-two sample size but the power function is not since $n_2(x_1)$ and $c_2(x_1)$ are mapped through the non-linear binomial distribution function. The problem is thus an integer non-linear program (INLP) (Lee *et al.*, 2011). Generic large-scale non-linear integer solvers to address this class of problems are mostly available under commercial or otherwise restrictive licenses (see, for instance, Artelys Knitro (Byrd *et al.*, 2006) or SCIP (Gleixner *et al.*, 2018)). Only recently, efforts to promote more open solutions are becoming available but have not yet reached production level stability (e.g. the Juniper solver (Kröger *et al.*, 2018)). Another disadvantage of generic INLP solvers is that they only produce *local* solutions since they guide their search using the gradients of the relaxed objective function and constraints. In the above given formulation, it remains unclear whether the problem is globally unimodular and thus amenable to such gradient-based optimisation approaches.

An alternative approach is to transform the problem to make it suitable for (mixed) integer *linear* programming (MILP) solvers. The advantage lies in the fact that tried and tested open MILP solvers are readily available (e.g., GLPK (GNU Project, 2020) or Cbc (COIN-OR initiative, 2020)). Furthermore, MILP problems can be solved to *global* optimality although the underlying integer optimisation problem is still NP-hard (Lee *et al.*, 2011). In principle, modern MI(N)LP solvers still rely on a Branch & Bound strategy. The difference to the custom implementation of Englert *et al.* (2013) is the fact that the search is guided by heuristics based on the relaxed problem where all variables are allowed to take real values. For linear objectives and constraints, this relaxed problem can be solved *globally* via the simplex algorithm (Wolsey *et al.*, 1999; Lee *et al.*, 2011). A formulation in terms of an integer linear program thus allows the use of more widely available solvers, exploits more effective implementations of the crucial Branch & Bound method, and guarantees a globally optimal and exact solution.

The challenge of formulating an effective exact and flexible solution method for problem (2.14)-(2.19) lies in its transformation to an integer *linear* program. The basic idea is similar to the concept of spline functions in regression. A spline function is given as a linear combination over a set of (potentially non-linear) basis functions. This can be exploited to implement non-linear regression via a spline basis using conventional effective methods for linear regression (Eubank, 1988). Note that the use of a sufficiently expressive set of basis functions may increase the dimensionality of the problem substantially. This means that a non-linear low dimensional problem can be transformed to a linear but high-dimensional problem.

In the following, a detailed approach to formulating problem 2.14)-(2.19) as mixed integer linear problem using auxiliary variables is described. The method is based on ideas developed in (Kunzmann *et al.*, 2016). Let

$$Y = \left(y[\,n_1, x_1, n_2, c_2\,] \right)_{\substack{n_1=n_1^{min},\ldots,n_1^{max} \\ x_1=0,\ldots,n_1 \\ n_2=0,\ldots,n_{\max}-n_1 \\ c_2=-\infty,0,\ldots,n_2-1,\infty}} \qquad y[\,n_1, x_1, n_2, c_2\,] \in \{0,1\} \tag{2.20}$$

be a sparse array of binary auxiliary variables indexed by the possible values of the respective parameters of the co-domains of the functions n_2 and c_2. The goal is to impose constraints on Y such that $y[\,n_1^*, x_1, n_2^*, c_2^*\,] = 1$ if and only if $n_1 = n_1^*$, $n_2(x_1) = n_2^*$ and $c_2(x_1) = c_2^*$. I.e., the $y[\,n_1^*, x_1, n_2^*, c_2^*\,]$ will act as 'selector' variables selecting for given n_1^* and x_1 the corresponding values n_2^* and c_2^*. First, one needs to make sure that only a single n_1^* is selected. This can be achieved by introducing additional binary auxiliary variables $z_{n_1} \in \{0,1\}, n_1 = n_1^{min}, \ldots, n_1^{max}$ and the following constraints

$$1 = \sum_{n_1} z_{n_1} \tag{2.21}$$

$$2\,n_{\max}\,z_{n_1} \geq \sum_{x_1,n_2,c_2} y[\,n_1, x_1, n_2, c_2\,] \quad \forall\ n_1 = n_1^{min}, \ldots, n_1^{max} \tag{2.22}$$

$$\sum_{n_2,c_2} y[\,n_1, x_1, n_2, c_2\,] = \sum_{n_2,c_2} y[\,n_1, x_1 - 1, n_2, c_2\,] \quad \forall\ \substack{n_1=n_1^{min}\ldots,n_1^{max} \\ x_1=1,\ldots,n_1}\ . \tag{2.23}$$

Here, the summation is over the entire range of the indices indicated under the summation symbol if not otherwise indicated. Constraint (2.21) implies that only one of the indicator variables z_{n_1} can be 1 giving it its meaning as a logical XOR. Constraint (2.22) guarantees that $z_{n_1} = 1$ if at least one configuration $n_1, n_2(x_1), c_2(x_1)$ is selected. Finally, constraint (2.23) ensures that n_1 is constant as a function of x_1. Jointly, these constraints give the variables z_{n_1} their meaning as selection variables for the stage-one sample size.

Next, it needs to be made sure that for any n_1 and any $x_1 = 0, \ldots, n_1$, both $n_2(x_1)$ and $c_2(x_1)$ are unique. This can be achieved by imposing

$$z_{n_1} = \sum_{x_1, n_2, c_2} y[\, n_1, x_1, n_2, c_2 \,] \quad \forall \; n_1 = n_1^{min}, \ldots, n_1^{max} \tag{2.24}$$

since z_{n_1} is 1 if and only if the stage one sample size is n_1.

Let $f(n_1, x_1, n_2, c_2)$ be any real-valued function. Its expected value with respect to X_1 given $n_1^*, n_2^*(\cdot)$, and $c_2^*(\cdot)$ as encoded in a configuration of Y can then be written as

$$\begin{aligned} &\mathrm{E}_p \left[\, f(n_1^*,\, X_1,\, n_2^*(X_1),\, c_2^*(X_1)) \,\right] \\ &= \sum_{n_1, x_1, n_2, c_2} \Pr_p[\, X_1 = x_1 \,] \cdot f(n_1, x_1, n_2, c_2) \cdot y[\, n_1, x_1, n_2, c_2 \,] \,. \end{aligned} \tag{2.25}$$

For instance, the expected sample size of a design $\mathcal{D}$ represented as an array Y is given by

$$\mathrm{E}_p[\, n^*(X_1) \,] = \sum_{n_1, x_1, n_2, c_2} \Pr_p[\, X_1 = x_1 \,] \, (n_1 + n_2) \, y[\, n_1, x_1, n_2, c_2 \,] \,. \tag{2.26}$$

Similarly, the power at response probability p is

$$\begin{aligned} &\Pr_p[\, X_2 > c_2^*(X_1) \,] \\ &= \sum_{n_1, x_1, n_2, c_2} \Pr_p[\, X_1 = x_1 \,] \, \Pr_p[\, X_2 > c_2(x_1) \,] \, y[\, n_1, x_1, n_2, c_2 \,] \,. \end{aligned} \tag{2.27}$$

Since both expressions are linear in Y, as are all constraints on Y, problem (2.14)-(2.19) is indeed an integer *linear* program in Y and $\{z_{n_1} \mid n_1 = n_1^{min}, \ldots, n_1^{max}\}$ and a global exact solution can be found using standard ILP solvers.

2.3.1. Choice of maximal sample size and additional constraints

Section 2.3 describes a general approach to formulating optimisation problems for generic single-arm two-stage designs with binary endpoint as integer linear programs. A naïve implementation would still be ineffective (although not impossible) due to the large number of possible configurations. To give an impression of the scale of the problems, consider $n_{\text{max}} = 100, n_1^{min} = 1$ and $n_1^{max} = n_{\text{max}}$. In this case, the number of weights is $|\, Y \,| = 4\,426\,525$. In practice, this number can be substantially reduced by imposing additional restrictions on the feasible space. This approach

is similar to the idea of considering only solutions with monotone conditional error function put forward by Englert *et al.* (2013) but entirely optional. Even without further constraints, a solution for typical problem sizes ($n_{\text{max}} < 100$) is still possible despite a needlessly long runtime.

The maximal overall sample size for the search space, n_{max}, can be chosen as a fixed multiple of the sample size of the standard binomial test. Throughout this thesis, a fixed multiple of 2 is used which is substantially larger than the maximal sample size considered by Englert *et al.* (2013) (+10% as compared to the maximal sample size in Simon's optimal design). The approximate sample size formula for the single-stage binomial test in Equation (1.6) is used to derive this upper bound.

There are two main reasons to restrict the feasible space. First and foremost speed. Due to the large number of variables and the need to pre-compute the coefficients for the problem can become a computational bottleneck and require more time than the actual solution of the problem. Luckily, some configurations can be ruled out *a priori* and the size of the feasible space can be reduced substantially. For many problems, e.g, searching over $0.2\, n_{\text{max}} \leq n_1 \leq 0.51\, n_{\text{max}}$ is sufficient to find the global optimum. Note that since n_{max} is chosen as twice the required sample size of a one-stage test $0.51\, n_{\text{max}}$ is still larger than the sample size derived from Equation (1.6) and the one-stage is binomial test is thus contained within the feasible space.

Secondly, an *a priori* reduction of the feasible space is also a means to impose additional regularity conditions on the optimal solution. For instance, the operational effort of conducting a very small second stage as compared to n_1 is rarely justifiable in practice. This can be prevented by imposing the restriction

$$|c_2(x_1)| < \infty \Rightarrow n_2(x_1) \geq \max(5, 0.1 \cdot n_1) \tag{2.28}$$

which sets the minimum size for stage two to the maximum of 5 or 10% of the stage-one sample size. Similarly, a very large second stage as compared to the first stage is rarely acceptable since it makes the overall duration and cost of a trial hard to calculate. E.g., imposing

$$|c_2(x_1)| < \infty \Rightarrow n_2(x_1) \leq 3 \cdot n_1 \tag{2.29}$$

restricts the relative increase to three times the stage-one sample size.

A specific challenge of optimal two-stage designs for single-arm trials with binary endpoint is the discreteness of the underlying test statistic. In some situations, this can lead to optimal solutions with non unimodal sample size functions (Englert *et al.*, 2013; Kunzmann *et al.*, 2016). Shan *et al.* (2016) addressed this implicitly by imposing a constraint that forced $n_2(\cdot)$ to be monotonically decreasing on the continuation region of the design. However, a monotonicity constraint is restrictive and it is not *a priori* clear whether optimal sample size functions tend to be increasing or decreasing in x_1 (see also Section 8.1). A less restrictive shape constraint is to only require unimodality of n_2 (Kunzmann *et al.*, 2016). In the majority of situations, optimal designs exhibit a unimodal sample size function. Global unimodality of the sample size function can be implemented by introducing further binary auxiliary indicator variables $u_{x_1} \in \{0, 1\}, x_1 = 0, \ldots, n_1$ and constraint sets for all n_1 and all

$x_1 = 0, \ldots, n_1$ (Kunzmann *et al.*, 2016)

$$\sum_{n_2', c_2'} n_2' \, y[\, n_1, x_1', n_2', \, c_2' \,] - \sum_{n_2'', c_2''} n_2'' \, y[\, n_1, x_1' - 1, n_2'', \, c_2'' \,] \geq n_{\max} \, (u_{x_1} - 1)$$

$$\forall \, x_1' < x_1 \tag{2.30}$$

$$\sum_{n_2', c_2'} n_2' \, y[\, n_1, x_1', n_2', c_2' \,] - \sum_{n_2'', c_2''} n_2'' \, y[\, n_1, x_1' - 1, n_2'', c_2'' \,] \leq n_{\max} \, (1 - u_{x_1})$$

$$\forall \, x_1' > x_1 \, . \tag{2.31}$$

These constraints are constructed such that they are only binding at a particular x_1 if $u_{x_1} = 1$. In that case, the constraints (2.30) ensure that all increments before x_1 are non-negative and the constraints of type (2.31) guarantee non-positive increments after x_1 thus making sure that x_1 is a global mode of the sample size function if $u_{x_1} = 1$. Finally, one needs to impose a constraint on the the minimum number of global modes (multiple global modes correspond to a locally constant sample size function) via

$$\sum_{x_1=0}^{n_1} u_{x_1} \geq 1 \, . \tag{2.32}$$

Note that multiple modes correspond to a locally constant sample size function, i.e., the maximal sample size is attained at neighbouring x_1. Jointly, these auxiliary variables and constraints ensure that the optimal solution has a (potentially non-unique) global mode. Since the addition of the global unimodality constraint is costly in terms of problem generation time and ILP solution time, it is only required as a backup option in case the initial solution without explicit unimodality constraints is not unimodal.

Further problem-specific considerations may be incorporated in the optimisation problem. Continuing a trial to stage-two with a low conditional power is rarely attractive for the sponsor of a clinical trial. The need for an adaptive recalculation due to low conditional power as described in Section 1.4.1 can be avoided by excluding all configurations that lead to low conditional power in the first place. Besides further reducing the feasible space, implementing a minimal conditional power already at the optimisation stage is evidently more effective than a heuristic *post hoc* recalculation. Similarly, most sponsors would probably prefer stopping for early efficacy over continuing to a second stage with more than 99% conditional power. A fairly conservative restriction would be to only allow a design to continue to stage-two if the conditional power given the interim result $X_1 = x_1$ lies between 50% and 99%. An in-depth example of how this methodology can be applied in practice is given in Section 8.1.

2.4. Other objective functions

Although the minimisation of expected sample size on the boundary of the null hypothesis is appealing in early clinical oncology due to the severe consequences of treating individuals with an ineffective new compound, this choice of objective function is by no means unequivocally accepted. The example discussed in Section 8.1 demonstrates that the choice of objective is even more important when considering generic two-stage designs than with less flexible group-sequential ones. For instance, the fact that the optimal generic two-stage design in the example situation has a slightly larger value of n_1 than Simon's group-sequential design indicates that the decision to consider expected sample size solely on the *boundary* of the null hypothesis may be inappropriate. One could argue, that the sample size on the interior of the null hypothesis is even more important in practice. Mander *et al.* (2010) argued that it might be more ethical to minimise sample size under the *alternative* hypothesis in situations with strong *a priori* evidence for the new treatments efficacy. Minimising expected sample size under the null hypothesis would lead to substantially prolonged trial duration in such cases. This, in turn, would prevent a larger patient collective from accessing a new treatment that is likely to be effective while the trial is still ongoing. These considerations give rise to the issue of how to incorporate performance over a range of response probabilities in the objective criterion and are discussed in more detail in Chapter 9.1.

Simon (1989) suggested to minimise the maximal sample size (and thus trial length) as an alternative objective criterion. This criterion is particularly attractive since it is independent of the response probability p. Minimax objectives are notoriously hard to optimise. Even in situations where all variables are continuous these problems suffer from the non-differentiability of the maximum function. In an ILP setting this issue can be overcome by the introduction of an additional continuous auxiliary variable at the cost of a substantially increased number of constraints. Let $\delta \in \mathbb{R}$ be that auxiliary variable with

$$\delta \geq \sum_{n_2, c_2} \big(n_1 + n_2 \big)\, y[\, n_1, x_1, n_2, c_2 \,] \quad \forall \; \substack{n_1 = n_1^{min}, \ldots, n_1^{max} \\ x_1 = 0, \ldots, n_1} \,. \tag{2.33}$$

Then, minimising δ corresponds to minimising the maximal sample size of the design.

Intuitively, it should be clear that a minimax objective favours less variable sample size functions. In fact, if the problem was smooth in all variables, the optimal solution would be a constant sample size. Only due to the discreteness of the underlying test statistic the optimal solution is not constant (Simon, 1989). Note that for generic two-stage designs for binary endpoints the minimiser under a minimax objective for the sample size function need not be not unique since there might be multiple designs fulfilling the error rate constraints and attaining the same minimal maximal sample size. This can only be overcome by combining the minimax criterion with another objective. As a practical solution one could minimise $(1 - \lambda)\, \delta + \lambda\, \mathrm{E}_p[\, n(X_1)\,]$ for $\lambda \in [\, 0, 1\,]$. The smallest possible maximal sample size is then found by identifying a non-unique solution to the problem for $\lambda = 0$ before iterative increasing λ to the largest $\lambda > 0$ under which the maximal sample size is still the same as for the solution

under $\lambda = 0$. For a comparison of different objective functions by means of a practical example, see Section 8.2.

3. Optimisation Under Uncertainty

This chapter is based on ideas previously discussed in Kunzmann *et al.* (2020a).

3.1. Assessing performance under uncertainty

The notion of 'optimality' crucially depends on the choice of objective criterion and different choices can lead to distinctly different 'optimal' designs (see Section 8.2). In particular, it the example discussed in Section 8.2 demonstrates that the characteristics of a design optimised for expected sample size under a single response probability p crucially depend on the choice of p. From a decision-theoretic perspective, minimising expected sample size is a rational choice since it corresponds closely with minimising trial-duration and the flexible costs of a trial. Bearing this in mind, a framework for incorporating uncertainty about the true value of p during planning is crucial. In this section, a general framework for scoring the performance of a design is introduced before going into the particulars of modelling uncertainty about the response probability p.

Let $\mathbb{D}$ be the space of feasible designs as defined in Sections 2.3 and 2.3.1. For any design $\mathcal{D} \in \mathbb{D}$, $\mathbb{X}_\mathcal{D} := \left\{ (x_1, x_2) \in \mathbb{Z}^2_{\geq 0} \,|\, x_1 = 0, 1, \ldots, n_1,\ x_2 = 0, 1, \ldots, n_2(x_1) \right\}$ is the corresponding sample space. A score function $s : \mathbb{D} \times \mathbb{Z}^2_{\geq 0} \times [\,0, 1\,] \to \mathbb{R}$ is a function mapping a design $\mathcal{D}$, a final trial outcome (x_1, x_2) and a response probability p to a numeric score value $s(\mathcal{D}, x_1, x_2, p)$. Without loss of generality, assume that lower scores are preferable and that $(x_1, x_2) \notin \mathbb{X}_\mathcal{D} \Rightarrow s(\mathcal{D}, x_1, x_2, p) = \infty$ for all $p \in [\,0, 1\,]$. A score function can thus be used to assign a quantitative value to each possible trial outcome under a given design and a given response probability. During the planning stage only the expected score

$$s(\mathcal{D}, p) := \mathrm{E}_p[\, s(\mathcal{D}, X_1, X_2, p)\,] \tag{3.1}$$

is relevant as the final outcome (X_1, X_2) is yet unknown[1]. Since $s(\mathcal{D}, p)$ depends on the unknown response probability p through the distribution of X_1 and X_2, assumptions about p need to be made in order to optimise the design. A consistent and principled way of doing so is by adopting a Bayesian perspective assuming that p itself is a random variable and follows a prior distribution. Without loss of generality, assume that the prior distribution of the response probability permits a density $\varphi(p)$ with respect to the Lebesgue measure on $[\,0, 1\,]$. Any valid functional with respect to φ of $s(\mathcal{D}, p)$ can then be used to assign an *unconditional* (on p) score value $s(\mathcal{D})$ to

[1] An extension to functional other than the expected value is possible but beyond the scope of this thesis.

a specific design. A natural choice is again taking the expected value with respect to the prior density, i.e.

$$s(\mathcal{D}) := \mathrm{E}_{\varphi(\cdot)}[\, s(\mathcal{D}, p)\,] = \int_0^1 s(\mathcal{D}, p)\, \varphi(p)\, \mathrm{d}p. \tag{3.2}$$

To make things more tangible, consider the example of the scoring a design by its final sample size, i.e.

$$s_{\mathrm{n}}(\mathcal{D}, x_1, x_2, p) := n(x_1)\ . \tag{3.3}$$

From this perspective, Simon's objective criterion of expected sample size on the boundary of the null hypothesis can be interpreted in two ways. The first corresponds to the prior being chosen as a point mass distribution on p_0, i.e., $\varphi(p) = \delta_{p_0}(p)$ (the Dirac-Delta distribution at p_0) and using the definition of equation (3.2). This interpretation raises the question why one would conduct a trial in the first place if one was already convinced that the true response rate is equal to p_0. The alternative interpretation is that the prior is unspecified with $\varphi(p_0) > 0$ and the unconditional score is defined in terms of the conditional expected value $\mathrm{E}_{\varphi(\cdot)}[\, s_{\mathrm{n}}(\mathcal{D}, p) \,|\, p = p_0\,]$. The latter view is consistent although it again highlights the fact that optimising a score under conditioning on a single response probability completely neglects the performance on a wide range of important response probabilities - in particular the interior of the null hypothesis $p < p_0$. A slight modification of Simon's objective criterion to

$$s_{n\,|\,p \leq p_0}(\mathcal{D}) := \mathrm{E}_{\varphi(\cdot)}[\, s_{\mathrm{n}}(\mathcal{D}, p) \,|\, p \leq p_0\,] \tag{3.4}$$

would already address this issue by weighting the expected sample size for all response probabilities withing the null hypothesis proportional to the prior density conditional on $p \leq p_0$. One reason for the fact that this modified criterion has not been discussed previously in the clinical trials literature is that $s_{n\,|\,p \leq p_0}(\mathcal{D})$ depends on the particular choice of φ instead of just the assumption that $\varphi(p_0) > 0$. This would require investigators to make their *a priori* assumptions explicit in a quantitative way by specifying φ. The general scepticism towards Bayesian methods in the clinical trial community might explain the past reluctance to accept such methodologies. Note, however, that the frequentist properties of the optimal design solely depend on the imposed constraints for its error rates (α and β) and are thus completely independent of the choice of objective criterion. Furthermore, there is no need to interpret φ in terms of a Bayesian prior density. The construction of the unconditional score function $s(\mathcal{D})$ as expected value with respect to $p \sim \varphi(\cdot)$ (see Equation (3.2)) can also be justified by interpreting φ as merely a normalised weight function on $[\,0, 1\,]$ since the Bayes theorem is not invoked at any point.

Informally, Chang *et al.* (1987) already introduced a similar concept by proposing to minimise the weighted sum of the expected value under the null and alternative hypothesis. Simon (1989) also considered Bayesian objective criteria as a potential future line of work. Recently, Jennison *et al.* (2015) and Pilz *et al.* (2019) discussed the minimisation of expected sample size averaged over a continuous prior in the setting of normally distributed endpoints.

More heuristically motivated score functions were proposed in the past. For instance, Liu *et al.* (2008) suggested a score that was used previously to compare two-stage designs (Kieser *et al.*, 2015). It evaluates a design at each p in relation to the corresponding single-stage design. Liu *et al.* (2008) also proposed to consider local averages to reflect uncertainty about the effect size. The score, however, is severely flawed in that it is not well-defined on the entire interval $[0, 1]$ and requires the specification of weight parameters that are hard to interpret in practice (Kunzmann *et al.*, 2020a).

In the following, the focus lies on the simpler to interpret expected sample size under the prior φ, i.e.

$$s_n(\mathcal{D}) = \int_0^1 s_n(\mathcal{D}, p)\, \varphi(p)\, \mathrm{d}p\,. \tag{3.5}$$

This score compromises between expected sample size under different response rates weighing them with their *a priori* relative likelihood under the chosen prior φ and can be seen as a direct extension of the minimisation of expected sample size under a single response probability p (e.g. Simon's objective criterion).

3.2. Prior choice

Any score taking into account the relative *a priori* likelihood of different response rates requires the choice of a prior density φ .

In the absence of any reliable *a priori* information, a non-informative prior seems to be a natural choice. Unfortunately, the notion of 'non-informativeness' is not unique. A first candidate could be the Jeffreys prior for a binomial experiment (Jeffreys, 1946). In one dimension, the Jeffreys prior coincides with the reference prior and is thus not only invariant under reparametrisation of the parameter space but also maximises, for any potential observation, the average Kullback-Leibler divergence between prior and posterior (Berger *et al.*, 2009). The latter property can be seen as a formalisation of the heuristic that a non-informative prior should give maximal weight to observed data. A disadvantage of the Jeffreys prior is its dependence on the sampling model, i.e. the design $\mathcal{D}$. Since the prior is supposed to be used in the definition of an objective criterion that can be used to elicit an optimal design, the prior itself cannot depend on the design. An alternative notion of non-informativeness is based on the idea of maximising the entropy of the prior on the parameter space. For the unit interval $[\,0, 1\,]$ this maximum entropy prior is the continuous uniform distribution (Park *et al.*, 2009). However, in many practical situations, a uniform prior on $[\,0, 1\,]$ cannot be seen as an adequate representation of *a priori* information despite its formal non-informativeness. Consider, for instance, the setting discussed in Section 1.3.1. With $p_0 = 0.2$ and $p_{\text{alt}} = 0.4$, the *a priori* chances of a very large response rate ($p \geq 0.5 > p_{\text{alt}}$) would be 50% under a uniform prior. In a situation where the trial is only powered for an alternative response rate of 0.4 this is clearly unrealistic bearing in mind that excessively large improvements over p_0 are rare in early clinical oncology (Ivanova *et al.*, 2016).

As a consequence, an informative or 'subjective' prior is often more adequate. It is important to stress that even under a subjective prior the frequentist properties of an optimal design only depend on the constraints on the maximal type one error rate and the type two error rate. A natural choice for the prior class are Beta distributions since they are conjugate to the binomial distribution. This means that the posterior after observing $X_1 = x_1$ or $X = x$ under a Beta prior is again a Beta distribution and available in closed form [2]. Let $\varphi_{a,b}(\cdot)$ be the density of a $\mathrm{Beta}(a, b)$ distribution. In a situation without any valid response data under the new treatment, the parameters a and b must be chosen based on a subjective assessment of the situation and potential prior biological evidence. To facilitate this, it might be easier to reparametrise the Beta distribution in terms of its mean μ and standard deviation σ. Both are rational expressions in a and b and solving the following system of equations algebraically

$$\mu = \frac{a}{a+b} \tag{3.6}$$

$$\sigma = \sqrt{\frac{a\,b}{(a+b)^2(a+b+1)}} \tag{3.7}$$

results in

$$a = \frac{-\mu^3 + \mu^2 - \mu\,\sigma^2}{\sigma^2} \tag{3.8}$$

$$b = \frac{(\mu - 1)\,(\mu^2 - \mu + \sigma^2)}{\sigma^2}\,. \tag{3.9}$$

Here it is assumed that $0 < \mu < 1$ and $0 < \sigma < 0.5$ since the standard deviation of a Beta distribution can never exceed 0.5.

Clinical oncology differs from most other drug development fields since phase I dose finding trials are mostly conducted in patients as well. This is due to the typically high toxicity of the tested compounds making trials in healthy subject difficult to justify. Consequently, there is often initial response data available from an earlier phase I during the planning stage of a phase II trial. Assume that x_0 responses out of n_0 subjects were recorded in phase I. The phase I data can then be incorporated in a subjective prior or a non-informative prior by Bayesian updating. This means that the phase I posterior becomes the prior for phase II. Conveniently, $\mathrm{Beta}(1, 1) = \mathrm{Uniform}(\,[\,0, 1\,]\,)$ holds which means that the phase I posterior is $\mathrm{Beta}(a + x_0, b + n_0 - x_0)$ if a $\mathrm{Beta}(a, b)$ prior (informative or not) was used to begin with.

The so-constructed prior might be further refined. Simply updating the prior for phase II with data from another, early phase trial implicitly assumes that the data between both trials is exchangeable. Differences in treatment regimes etc. often make this assumption implausible. Several suggestions of how to combine trial data with 'historic' data are put forward in the literature (Viele *et al.*, 2014). Due to the small sample sizes in phase I/II, only the simplest of methods for adjusting the impact of

[2] This holds despite X not being marginally distributed according to a binomial distribution. It is sufficient that that $X_2 \mid X_1 = x_1, n_2(x_1) > 0$ does follow a binomial distribution.

historical data on the prior are feasible. An attractive and hands-on approach that is applicable both in cases where the phase I data cannot be assumed to be exchangeable is robustification. The core insight is that priors with heavy tails react much more sensitively to situations where the data is not well explained by the prior. In such a situation, the posterior of a robust prior fits the observed data much closer than under a non-heavy tailed prior. Heavy tails can be achieved by defining a mixture prior with the original prior and a uniform component (Schmidli *et al.*, 2014). Since the informative prior is a Beta distribution and the uniform distribution on $[\,0, 1\,]$ is also a Beta distribution, the resulting two-component mixture is a mixture of conjugate priors and the posterior can be computed analytically per component. I.e., the density of the robust informative prior is defined as

$$\varphi_{a,b,\epsilon}(p) := (1-\epsilon)\,\varphi_{a,b}(p) + \epsilon\,\varphi_{1,1}(p). \tag{3.10}$$

The value of ϵ represents a confidence for the overall credibility of the informative prior for the new phase II trial. Technically, the effect of a robustification with a uniform density on the resulting unconditional score can be explained by studying the functional derivative of an unconditional score with respect to local changes at a response probability p. The functional derivative of a functional F at f is defined as

$$\frac{\partial F}{\partial f} := \lim_{\epsilon \to 0} \frac{F[\,f + \epsilon\,\phi\,] - F[\,f\,]}{\epsilon} \tag{3.11}$$

(Giaquinta *et al.*, 1996) and captures the change in the functional F due to local variations analogously to the derivative of a function. The explicit form for the functional derivative of functionals of the form

$$F[\,f\,] := \int_a^b L(x, f(x), f'(x))\ \mathrm{d}x \tag{3.12}$$

with a twice differentiable function L evaluated at x is

$$L_2\big(x, f(x), f'(x)\big) - \frac{\partial}{\partial x} L_3\big(x, f(x), f'(x)\big) \tag{3.13}$$

where $L_2(x, f(x), f'(x))$ is the partial derivative of L with respect to its second argument and $L_3((x, f(x), f'(x)))$ the partial derivative of L with respect to its third argument (Giaquinta *et al.*, 1996). A score $s(\mathcal{D})$ with prior $\varphi_{a,b,\epsilon}$ can be expressed as a functional of this particular form when setting $L(p, s(\mathcal{D}, p), \partial s(\mathcal{D}, p)/\partial p) = \varphi_{a,b,\epsilon}(p)\ s(\mathcal{D}, p)$ (see equation (3.2)). Since this particular inner function L does not depend on the first derivative of the score function, the second term in (3.13) vanishes and the functional derivative with respect to score changes at a response probability p simplifies to

$$\frac{\partial s(\mathcal{D})}{\partial s(\mathcal{D}, \cdot)}(p) = \varphi_{a,b,\epsilon}(p) \geq \epsilon\,. \tag{3.14}$$

This means that under a robust prior the contribution of the conditional (on p) score $s(\mathcal{D}, p)$ to the unconditional score $s(\mathcal{D})$ is bounded from below to ϵ. Without the

robustification $\varphi_{a,b,}(p)$ can go to zero quickly in the tails implying that changes in the conditional score in the tail regions of the prior do not affect the overall score.

In a last step, it might be necessary to make sure that the robust prior does not put too much weight on completely implausibly large response rates. This is particularly important when the anticipated response rates are low and most of the informative prior's mass is concentrated on low values of p. The simplest solution is to condition the robust prior on $p \leq \overline{p}$ where $\overline{p}$ is the maximal plausible response rate under the new treatment:

$$\varphi_{a,b,\epsilon}|_{\leq \overline{p}}(p) := \frac{\mathbf{1}_{p \leq \overline{p}}(p)\, \varphi_{a,b,\epsilon}(p)}{\int_0^{\overline{p}} \varphi_{a,b,\epsilon}(q)\ \mathrm{d}q}. \tag{3.15}$$

Since the order of conditioning is commutative, the posterior under such a restricted Beta mixture can still be computed analytically by first updating the individual components before conditioning on $p \leq \overline{p}$ again. The process of constructing a prior density along the lines described above is illustrated in Section 9.1.

3.3. Power constraints under uncertainty

So far, only the consequences of incorporating uncertainty in the objective function (expected sample size) were discussed. The error rate constraints remained unchanged. This is sensible for strict type one error rate control on the null hypothesis since the constraint is independent of any prior assumptions. To control the maximal type one error rate on the null hypothesis, it is sufficient to impose a constraint on the maximal 'power' at the boundary of $\mathcal{H}_0$ if the designs power function is monotone in p. Monotonicity of the power function is not explicitly implemented in the optimisation problem. In practice, however, monotonicity of the power function can easily be verified after the optimisation problem was solved. Consequently, the constraint on the maximal power at the boundary of the null hypothesis remains unaffected by considerations about the *a priori* relative likelihood of different response rates that is encoded in the prior density. However, this is not the case for the classical power constraint

$$\Pr_{p_{\text{alt}}}[\, X_2 > c_2(X_1)\,] \geq 1 - \beta\ . \tag{3.16}$$

The validity of this constraint critically depends on the justification of the alternative response rate p_{alt}. Ideally, p_{alt} would be chosen as the smallest still clinically relevant response rate. Early oncological trials are often powered for a point alternative with an absolute rate difference of 0.2 over p_0 (Ivanova *et al.*, 2016). This is partly due to the fact that a sufficiently large benefit in terms of the surrogate endpoint response rate is generally deemed necessary to show efficacy in terms of overall survival in a subsequent phase III trial (Simon, 1989; Ivanova *et al.*, 2016). Often, however, it must be assumed that the relatively optimistic choice of $p_{\text{alt}} = p_0 + 0.2$ is more driven by the desire to obtain a feasible sample size. In some cases, response probabilities as low as $p_0 + 0.1$ may still be considered clinically relevant. The sample size required to reliably detect such small differences, however, is too large for early phase II studies. For

instance, to detect a rate difference of 0.1 for $p_0 = 0.2$ with $\alpha = 5\%$ and 80% power the required sample size of $n = 130$ (see Equation (1.6)) is more than twice as large as the typical sample sizes reported (Ivanova *et al.*, 2016). In situations like these, the rationale for the value of p_{alt} is more driven by *a priori* likelihood arguments: If the true effect can be assumed to be greater than the minimal clinically relevant one, it is reasonable to power for a larger, likely effect and risk a failed trial for smaller but still relevant response rates. Let p_{MCR} be the minimal clinically relevant response rate under the new treatment. The problem is then to pick $p_{\text{alt}} > p_{\text{MCR}}$ such that prior evidence for $p > p_{\text{MCR}}$ can be exploited in a principled way without risking an underpowered study. To this end it is again assumed that a prior density $p \sim \varphi(\cdot)$ is available.

Taking a Bayesian perspective, it is sensible to choose p_{alt} such that the *a priori* probability of exceeding the target power of $1 - \beta$ is larger than or equal to a confidence level γ. Let $\text{power}(p) := \Pr_p[\, X_2 > c_2(X_1)\,]$ be the power function of the yet unspecified design. Since p is modelled as a random variable under the Bayesian paradigm, $\text{power}(p)$ is a random variable. A high power is only desirable for relevant effect sizes. It is thus sensible to determine the size of a trial such that the probability of exceeding a power of $1-\beta$ exceeds a defined threshold γ given that there is indeed a relevant effect. Assuming that the power function is monotone

$$\Pr_{\varphi(\cdot)}[\, \text{power}(p) \geq 1 - \beta \,|\, p \geq p_{\text{MCR}}\,] = \gamma \tag{3.17}$$

$$\Leftrightarrow\ \text{power}(q_{1-\gamma}) \geq 1 - \beta \tag{3.18}$$

holds, where $q_{1-\gamma}$ is the $(1-\gamma)$-quantile of the prior $\varphi(\cdot)$ conditional on $p > p_{\text{MCR}}$ (Kunzmann *et al.*, 2020a). Conditioning on $p \geq p_{\text{MCR}}$ ensures that only the power to reject $\mathcal{H}_0$ under actually relevant response rates is taken into account. The advantage of this construction lies in the fact that the power constraint technically remains a constraint on a single point of the power curve - only a novel justification of $p_{\text{alt}} = q_{1-\gamma} \geq p_{\text{MCR}}$ is given. Under this prior quantile approach the response probability at which the power constraint is imposed now depends on φ through the quantile function of the conditional prior. A disadvantage of the approach is that it requires yet another parameter, γ, whose role can easily be confused with that of $1 - \beta$.

Alternatively, and similar to the definition of $s_n(\mathcal{D})$ in Section 3.1, expected power

$$s_{\text{power}}(\mathcal{D}) := \text{E}_{\varphi(\cdot)}\left[\, \Pr_p[\, X_2 > c_2(X_1)\,] \;|\, p \geq p_{\text{MCR}}\,\right] \tag{3.19}$$

$$= \int_{p_{\text{MCR}}}^{1} \Pr_p[\, X_2 > c_2(X_1)\,] \;\frac{\varphi(p)}{\Pr_{\varphi(\cdot)}[\, p \geq p_{\text{MCR}}\,]}\; \text{d}p \tag{3.20}$$

can be used as a functional of the power curve. This expected score is generated by

$$s_{\text{power}}(\mathcal{D}, x_1, x_2, p) := \frac{\mathbf{1}_{x_2 > c_2(x_1) \wedge p \geq p_{\text{MCR}}}(x_1, x_2)}{\Pr_{\varphi(\cdot)}[\, p \geq p_{\text{MCR}}\,]}\,. \tag{3.21}$$

For $p_{\text{MCR}} = p_0$ this definition coincides with the definition of expected power given by Brown *et al.* (1987) in the context of normally distributed test statistics. Since not all $p > p_0$ might be considered relevant enough to warrant further investigation

of a new therapy in phase III, the distinction between p_{MCR} and the boundary of the null hypothesis is particularly important in the planning of phase II trials for oncology. The advantage of using expected power over the prior quantile approach is that no further parameters besides the already chosen prior and p_{MCR} need to be specified. Furthermore, there is a direct connection between expected power and the success probability of a trial that is explored in more detail in Section 3.4.

A constraint on expected power is qualitatively different from constraining the power on a single response probability p_{alt} - no matter how p_{alt} is chosen. To see this, the relative contribution of power at different response rates on the function of the power curve can be studied. In the case of putting a constraint on power at p_{alt}, by definition, the power at $p \neq p_{\text{alt}}$ is irrelevant to the fulfilment of the constraint. This can be formalised by the fact that the total differential of power at p_{alt} with respect to changes in the power curve at p_{alt} and $p' \neq p_{\text{alt}}$ only depends on changes of power at p_{alt}, i.e.

$$\begin{aligned} \mathrm{d}\,\mathrm{power}(p_{\text{alt}}) &= \frac{\partial\,\mathrm{power}(p_{\text{alt}})}{\partial\,\mathrm{power}(p_{\text{alt}})}\,\mathrm{d}\,\mathrm{power}(p_{\text{alt}}) + \frac{\partial\,\mathrm{power}(p_{\text{alt}})}{\partial\,\mathrm{power}(p')}\,\mathrm{d}\,\mathrm{power}(p') && (3.22)\\ &= \mathrm{d}\,\mathrm{power}(p_{\text{alt}})\,. && (3.23) \end{aligned}$$

Consequently, during optimisation only changes to the design that affect power at p_{alt} can help fulfilling the power constraint. The power curve at $p \neq p_{\text{alt}}$ is completely irrelevant to the optimal solution. For expected power, however, the total differential with respect to changes in the power curve at two response probabilities $p_0 \leq p_1 \leq p_2 \leq \overline{p}$ is given by

$$\mathrm{d}\,s_{\text{power}}(\mathcal{D}) = \frac{\partial s_{\text{power}}(\mathcal{D})}{\partial s_{\text{power}}(\mathcal{D}, p_1)}\,\mathrm{d}\,s_{\text{power}}(\mathcal{D}, p_1) + \frac{\partial s_{\text{power}}(\mathcal{D})}{\partial s_{\text{power}}(\mathcal{D}, p_2)}\,\mathrm{d}\,s_{\text{power}}(\mathcal{D}, p_2)\,. \tag{3.24}$$

To asses the relative contribution of the values of the power curve at p_1 as compared to p_2, it is insightful to ask how much power at p_1 would have to change in order to offset a change of power at p_2. Thus setting the total differential to zero and solving for $\mathrm{d}\,s_{\text{power}}(\mathcal{D}, p_2)$ yields

$$\begin{aligned} \mathrm{d}\,s_{\text{power}}(\mathcal{D}, p_2) &= -\frac{\dfrac{\partial s_{\text{power}}(\mathcal{D})}{\partial s_{\text{power}}(\mathcal{D}, p_1)}}{\dfrac{\partial s_{\text{power}}(\mathcal{D})}{\partial s_{\text{power}}(\mathcal{D}, p_2)}}\,\mathrm{d}\,s_{\text{power}}(\mathcal{D}, p_1) && (3.25)\\ &= -\frac{\varphi(p_1)}{\varphi(p_2)}\,\mathrm{d}\,s_{\text{power}}(\mathcal{D}, p_1). && (3.26) \end{aligned}$$

The relative importance thus directly corresponds to the relative likelihood of the response rates p_1 and p_2 under the prior density. To keep overall expected power constant, a decrease of $0.01 = 1\%$ in power at p_1 ($\mathrm{d}\,s_{\text{power}}(\mathcal{D}, p_1) = -0.01$)) can be offset by increasing power at p_2 by $\mathrm{d}\,s_{\text{power}}(\mathcal{D}, p_2) = 0.01\,\varphi(p_1)/\varphi(p_2)$. Firstly, this argument demonstrates formally, that indeed the entire power curve for $p_{\text{MCR}} \leq p \leq \overline{p}$

contributes towards the fulfilment of a constraint on expected power during optimisation. Secondly, it illustrates the implicit trade-off between power at different response probabilities when constraining expected power instead of power at a single point-alternative. For a worked example of how Bayesian power constraints affect the optimal solution see Section 9.2.

3.4. A utility-based approach

From the perspective of a pharmaceutical company's shareholders, the utility of a trial is mainly given by the expected future financial payout. In phase II, the future payout is notoriously hard to judge since a new compound still needs to successfully complete phase III before being filed for approval with regulatory agencies. Any revenues can only be generated after an approval has been issued. Here, two relevant cases need to be distinguished. Firstly, if the phase II trial rejects the null hypothesis and the response rate is indeed relevant, the expected payout depends on the chances of the new compound making it all the way through phase III and to market. The expected payout in this case thus needs to make assumptions about the market potential and the chances of the compound to complete a phase III programme given that $p \geq p_{\text{MCR}}$. Secondly, if the trial rejects the null hypothesis but the true response rate is *not* relevant, the compound will still continue to phase III. Eventually, however, the new drug will fail to demonstrate efficacy and will not make it to market[3]. The expected payout in this case is thus generally negative since phase III studies are still conducted but there is never any revenue generated from the drug. These considerations can be formalised within the score-framework discussed in Section 3.1.

Let $\lambda_{+|+}$ be the future payout for a true positive phase II finding and $\lambda_{+|-}$ the future cost for a false positive finding. For sake of simplicity, assume that both $\lambda_{+|-}$ and $\lambda_{+|+}$ are defined on the scale of the average marginal[4] per-patient cost within the planned phase II trial. I.e., if the average marginal per-patient cost was 50 000 US$ and $\lambda_{+|+} = 200$, then the future payout upon successful rejection of the null and a truly relevant response probability was 10 million US$. Let

$$\begin{aligned} \mathrm{s}_u(\mathcal{D}, x_1, x_2, p) := \lambda_{+|+} \, \mathbf{1}_{x_2 > c_2(x_1) \wedge p \geq p_{\text{MCR}}}(x_1, x_2, p) \\ + \lambda_{+|-} \, \mathbf{1}_{x_2 > c_2(x_1) \wedge p < p_{\text{MCR}}}(x_1, x_2, p) \\ - n(x_1) \end{aligned} \tag{3.27}$$

to formalise the above line of arguments. Following the same pattern as in Section 3.1

[3] Here it is tacitly assumed that the joint probability of a type one error in phase II and phase III is negligible to keep things simple. The general line of argument is unaffected by a more complex future-payout model though.

[4] The marginal costs ignore any fixed costs of the trial.

the expected score for given response probability p is

$$\begin{aligned} \mathrm{s}_u(\mathcal{D}, p) := \lambda_{+|+} \, \Pr_p[\, X_2 > c_2(X_1) \,] \, \mathbf{1}_{p \geq p_{\text{MCR}}} \\ - \, \lambda_{+|-} \, \Pr_p[\, X_2 > c_2(X_1) \,] \, \mathbf{1}_{p < p_{\text{MCR}}} \\ - \, \mathrm{E}_p[\, n(X_1) \,] \end{aligned} \tag{3.28}$$

and the unconditional expected utility score with respect to the prior φ is

$$\begin{aligned} \mathrm{s}_u(\mathcal{D}) := \lambda_{+|+} \, \Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1), \, p \geq p_{\text{MCR}} \,] \\ - \lambda_{+|-} \, \Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1), \, p < p_{\text{MCR}} \,] \\ - \, \mathrm{E}_{\varphi(\cdot)}[\, n(X_1) \,] \, . \end{aligned} \tag{3.29}$$

The unconditional probability

$$\begin{aligned} & \Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1), p \geq p_{\text{MCR}} \,] \\ = \, & \Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1) \,|\, p \geq p_{\text{MCR}} \,] \, \Pr_{\varphi(\cdot)}[\, p \geq p_{\text{MCR}} \,] \end{aligned} \tag{3.30}$$

closely resembles the definition of 'probability of success' proposed by Spiegelhalter *et al.* (1986). The only difference is that Spiegelhalter implicitly assumed that $p_{\text{MCR}} = p_0$. Allowing the probability of success to depend on a minimally clinically relevant response rate is thus a slightly more flexible definition and will be used throughout the remainder of this thesis. Also, $\Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1) \,|\, p \geq p_{\text{MCR}} \,]$ is exactly the definition of expected power (cf. 3.19). Similarly, $\Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1), p < p_{\text{MCR}} \,]$ corresponds to the marginal probability of a type one error. Finally, $\mathrm{E}_{\varphi(\cdot)}[\, n(X_1) \,]$ is the overall expected sample size of the phase II trial.

Note that s_u does not incorporate the fixed costs of the phase II trial. Yet, adding a constant to s_u does affect the solution obtained from minimising the score and can therefore be omitted for sake of simplicity. The fixed costs are only relevant when a decision has to be reached as to whether a planned trial should actually be conducted. Whenever the expected utility s_u is less than the fixed costs, it would be rational to *not* conduct the trial.

Jennison *et al.* (2000, 2015) applied utility maximisation to the planning of clinical trial designs and proposed optimisation of a similar score in the setting of (asymptotically) normally distributed test statistics. Due to the typically larger sample sizes in such settings, they analysed a relaxed version of the problem ignoring the integer restriction on sample sizes via gradient-based methods. The implicit definition of their expected utility score

$$\int_{-\infty}^{\infty} \varphi(\theta) \big(\, \mathrm{power}(\theta) - \lambda \, \mathrm{E}_\theta[\, n(\theta) \,] \big) \, \mathrm{d}\theta, \tag{3.31}$$

where θ is the location parameter of interest and $\varphi(\theta)$ is the corresponding prior (equation (10), Section 6.1, Jennison *et al.* (2015)), however, differs in two crucial aspects. Firstly, by integrating power over all possible values without distinction of the

respective relevance they implicitly count false positives as successes. Consequentially a larger probability to reject the null is always favourable (strangely also under $p < p_0$). Secondly, as a direct consequence of this lack of differentiation between power on the null hypothesis and on the alternative, their score cannot be maximised directly since $n = 0$ and always rejecting the null hypothesis would be optimal. Instead, they need to explicitly incorporate a constraint on the maximal type one error rate to obtain sensible results.

With s_u, however, this is not necessary since the *expected* type one error rate is naturally penalised via $\lambda_{+|-}$. This highlights the fact that under s_u, any additional error rate constraints are optional. Rather, the error rates (i.e., the power curve) of the optimal design are determined naturally via the real-world consequences (payout) of the respective trial outcomes. Both α and β are thus implicitly determined in a situation-specific manner taking into account both future payout (utility) and *a priori* knowledge φ. Consequently, the maximal type one error rate of a utility-maximising design can be both lower or higher than the standard 5%. In cases where the maximal type one error rate is substantially higher than 10% (which is sometimes still accepted by regulators for early oncology trials (Simon, 1989)), one could simply impose a constraint on the maximal type one error rate as discussed in the preceding chapter. The main practical benefit of a utility-based approach to trial design is to identify situations where it is in the sponsors interest to exceed the regulatory minimal requirements on type one or type two error rates, i.e. where it is beneficial to have lower type one error rates or larger power than usual. An in-depth example on how utility-based methods can be implemented in practice is given in Section 9.3.

4. Bayesian Inference

In the preceding chapters of this thesis, the exclusive focus was on the derivation of optimal two-stage designs for single-arm trials with binary endpoint. After the conclusion of a trial, however, it is not only of interest whether the respective null hypothesis $\mathcal{H}_0 : p \leq p_0$ can be rejected. Beyond this binary result, a quantification of the post-trial evidence for individual response rates and against the null hypothesis are required to guide further decision making for a potential phase III. In Chapter 3, the importance of the correct and transparent specification of *a priori* assumptions on the relative plausibility of different values of the true unknown response probability p was discussed. The Bayesian framework offers a consistent and accessible way of incorporating these assumptions in a prior density φ. Under correctly specified constraints on the type one and type two error rates, the frequentist properties of the final test decision can be controlled exactly - despite the necessarily subjective choice of the prior (cf. Section 3.1). Since a prior φ should thus be specified during the planning phase of a trial, a natural framework for post-trial inference is the Bayesian one (Jeffreys, 1998). Bayesian inference is rarely used to analyse (confirmatory) clinical trials due to general concerns about biased estimates and frequentist properties of Bayesian credible intervals. This is particularly the case for pivotal phase III studies that intended to support approval by a regulatory body. A thorough treatment of the vast literature on the Bayesian-vs-Frequentist dispute is beyond the scope of this thesis. For an excellent historical overview of the matter, see Salsburg (2001). Instead, methods for applying both frameworks to the situation at hand are presented.

The contents of this and the following chapter are partly based on results discussed in Kunzmann *et al.* (2017a) and Kunzmann *et al.* (2017b).

4.1. Inference using the planning prior

Bayesian inference is exclusively based on the posterior distribution of the unknown parameter p given the observed data $(X_1, X_2) = (x_1, x_2)$. Since Bayesian inference is consistent with the likelihood principle, the sampling scheme (i.e. the design $\mathcal{D}$) under which the conditionally independent stage-wise observations (x_1, x_2) were obtained is irrelevant (Jeffreys, 1998). The posterior only depends on the data likelihood and the prior φ. Let

$$\varphi(p \,|\, x_1, x_2) := \frac{\Pr_p[\, X_1 = x_1, X_2 = x_2 \,]\, \varphi(p)}{\displaystyle\int_0^1 \Pr_p[\, X_1 = x_1, X_2 = x_2 \,]\, \varphi(p)\ \mathrm{d}p} \tag{4.1}$$

be the posterior density. For the proposed prior class of truncated mixtures of Beta distributions (cf. Section 3.2), the posterior can be computed analytically since the

Beta distribution is the conjugate prior for the binomial data distribution. The posteriors of a mixture distribution is the mixture of the individual component posteriors and therefore also directly available. Finally, the proposed optional truncation to control the upper tail behaviour of the prior is also a conditioning operation and interchangeable with conditioning on the data. Thus the posterior is available in closed form for a $\varphi_{a,b,\epsilon}|_{p\leq\overline{p}}$ prior (cf. Section 3.2). Let

$$f(p,a,b) := \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\,p^{a-1}(1-p)^{b-1} \tag{4.2}$$

be the probability density function of a Beta distribution with parameters a and b and $F(p,a,b)$ the corresponding cumulative probability function. Using $x = x_1 + x_2$ the posterior density is

$$\varphi_{a,b,\epsilon}|_{p\leq\overline{p}}(p \mid x_1, x_2) = \begin{cases} 0 \quad \textit{if}\, p > \overline{p} \\ \epsilon\,\dfrac{f(p, 1+x, 1+n(x_1)-x)}{F(\overline{p}, 1+x, 1+n(x_1)-x)} \\ \quad +(1-\epsilon)\,\dfrac{f(p, a+x, b+n(x_1)-x)}{F(\overline{p}, a+x, b+n(x_1)-x)} \quad \text{else .} \end{cases} \tag{4.3}$$

The posterior encodes rich information about the combined *a priori* and within-trial evidence. The entire distribution should be visually inspected before reducing it to functionals like the posterior-mean, a frequently used Bayesian point estimate, or the credible interval which is the Bayesian counterpart to the frequentist confidence interval. Here, the (1-α)-posterior credible interval is given by the closed interval delimited by the $\alpha/2$ and $1-\alpha/2$ quantiles.

A further advantage of drawing inference mainly from a Bayesian posterior probability lies in the fact that it can serve as prior for subsequent phase III studies in a similar way as φ was employed to plan the phase II trial. This is of particular interest when the phase II study simultaneously collects overall-survival data (the primary endpoint in phase III) and the surrogacy relationship between the two quantities is modelled (Rufibach *et al.*, 2020).

4.2. Objective Bayesian inference

One way of addressing the bias introduced by using the informative planning prior under the Bayesian paradigm is to compute the posterior under a non-informative prior. The concept of using a different 'analysis prior' for inference was put forward in (O'Hagan *et al.*, 2005). The rationale behind the choice of an 'analysis prior' is primarily guided by the frequentist properties of the resulting point estimator instead of trying to correctly quantify any *a priori* evidence. For inference, the design can be considered fixed and the theoretically attractive Jeffreys prior is a feasible choice. Bayesian analysis of binomial outcomes using Jeffreys priors is known to show favourable properties in terms of the frequentist coverage probabilities for the one-stage case (DasGupta *et al.*, 2001). The Jeffreys prior does depend on the sampling

scheme (i.e. the design) through its dependence on the Fisher information. Thus, for any particular design, the corresponding Jeffreys prior will typically be different from the Jeffreys prior of a single-stage binomial test where it is given by a Beta distribution with parameters $a = 0.5$ and $b = 0.5$ (Jeffreys, 1946; Kunzmann *et al.*, 2017a). To derive the Jeffreys prior for a particular two-stage design, one first needs to compute the corresponding Fisher information. The likelihood function $\ell(p \mid x_1, x_2)$ under any two-stage design $\mathcal{D}$ is proportional to $p^{x_1+x_2}(1-p)^{n(x_1)-x_1-x_2}$. This expression, in turn, is also proportional to the likelihood of a single-stage binomial experiment with $x = x_1 + x_2$ out of $n = n(x_1)$ responses and response probability p. This greatly simplifies the computation of the Fisher information

$$\mathcal{I}_{\mathcal{D}}(p) = \mathrm{E}_p\left[\left(\frac{\partial}{\partial p}\log\big(\ell(p \mid X_1, X_2)\big)\right)^2\right]. \tag{4.4}$$

Exploiting the proportionality, the derivative of the log likelihood function is given by

$$\frac{\partial}{\partial p}\log\big(\ell(p \mid x_1, x_2)\big) \tag{4.5}$$

$$= \frac{\partial}{\partial p}\left(\log\left(\binom{n_1}{x_1}\binom{n_2(x_1)}{x_2}\right) + x\log(p) + (n(x_1) - x)\log(1-p)\right) \tag{4.6}$$

$$= \frac{\partial}{\partial p}\big(x\log(p) + (n(x_1) - x)\log(1-p)\big) \tag{4.7}$$

$$= \frac{x}{p} - \frac{n(x_1) - x}{1-p}. \tag{4.8}$$

Thus, the Fisher information of $\mathcal{D}$ is

$$\mathcal{I}_{\mathcal{D}}(p) = \mathrm{E}_p\left[\left(\frac{X_1 + X_2}{p} - \frac{n(X_1) - X_1 - X_2}{1-p}\right)^2\right] \tag{4.9}$$

$$= \sum_{(x_1, x_2)\in\mathbb{X}_{\mathcal{D}}} \binom{n_1}{x_1}\binom{n(x_1)-n_1}{x_2} p^x (1-p)^{n(x_1)-x}\left(\frac{x}{p} - \frac{n(x_1)-x}{1-p}\right)^2 \tag{4.10}$$

where $\mathbb{X}_{\mathcal{D}}$ is the sample space corresponding to the design $\mathcal{D}$ and the Jeffreys prior $\varphi^{\mathcal{D}}_{\text{Jeffreys}}$ is proportional to the square root of the Fisher information

$$\varphi^{\mathcal{D}}_{\text{Jeffreys}}(p) \propto \sqrt{\mathcal{I}_{\mathcal{D}}(p)}\,. \tag{4.11}$$

Finally, the normalising constant can be evaluated by numerical integration. A worked example comparing the impact of the prior choice on inference is given in Chapter 10.

5. Frequentist Inference

The contents of this chapter are partly based on Kunzmann *et al.* (2017a) and Kunzmann *et al.* (2017b).

Besides the Bayesian one, frequentist inference is the other major paradigm in statistics. Here, the predominant concept is the avoidance of any form of 'subjectivity' or bias introduced by the adoption of a situation-specific prior distribution which is necessarily subjective. The frequentist paradigm is still dominant in the analysis of clinical trials and particularly so for confirmatory and pivotal phase III trials. To complement the Bayesian methods outlined in Chapter 4, this chapter develops an entirely frequentist approach to inference in single-arm two-stage designs for binary endpoints and contrasts it with the Bayesian one.

After the conclusion of a trial it is not only of interest whether the respective null hypothesis can be rejected but also to quantify the evidence against it and to give an estimate of the true response rate p. Usually, a p value together with a point estimate of the true response rate is provided together with an interval estimate, typically a confidence interval (see, e.g., U.S. Food and Drug Administration (1998)).

5.1. The one-stage design situation

First, consider a simple single-stage binomial test, with X out of n responses for $\mathcal{H}_0 : p \leq p_0$. Let the critical value c be chosen such that the one-sided significance level is α. The p value for $\mathcal{H}_0$ is defined as the probability of observing an outcome at least as extreme seen from the null hypothesis as the actually observed one given that p_0 is true (Fisher, 1925; Wasserstein *et al.*, 2016), i.e.,

$$\rho(x) := \Pr_{p_0}\left[X \succeq x \right]. \tag{5.1}$$

Here, the relation '$\succeq$' denotes an ordering on the outcome space $\{0, \ldots, n\} \subset \mathbb{Z}$ and the p value is denoted $\rho(x)$ to avoid confusion with the response probability p. Formally, an ordering on $\{0, \ldots, n\}$ is defined as a function

$$\succeq\colon \{0, \ldots, n\}^2 \to \{0, 1\} \tag{5.2}$$

$$x \succeq x' :\Leftrightarrow \succeq (x, x') = 1 \Leftrightarrow x \text{ is as extreme or more extreme than } x'. \tag{5.3}$$

In the one-stage case, a natural choice for $\succeq$ is given by the canonical ordering of the real numbers, i.e., $x \succeq x' \Leftrightarrow: x \geq x'$ since it is unequivocal that larger values of x provide more evidence against $\mathcal{H}_0$ than smaller ones. One way of justifying this intuition more formally is via a likelihood argument since $x \succeq x' \Leftrightarrow \ell(x, p_0) \leq \ell(x, p_0)$ where $\ell(x, p)$ is the binomial likelihood of the observation x given response

rate p. The choice of ordering can also be justified in terms of point estimation. Let $\widehat{p} : \{0, \dots, n\} \to [\,0, 1\,]$ be an estimator of the response probability p. In the setting considered here, it is reasonable to assume that any sensible estimator is monotone in x, i.e, $\widehat{p}(x) \geq \widehat{p}(x') \Leftrightarrow x \geq x'$. The natural ordering is thus also the same as the one induced by ordering outcomes by their point estimates. If one were to come up with any different ordering $\succeq' \neq \geq$ on $\{0, \dots, n\}$ in terms of extremeness from $\mathcal{H}_0$, this would immediately imply the existence of a pair of observations x, x' with $x \succeq' x'$ but $x \not\geq x'$, i.e, $x < x'$. This would mean that, under $\succeq'$, there was at least one situation in which a smaller number of responses, x, was considered more extreme from $\mathcal{H}_0$ than a large number of responses, x'. Consequently, due to the assumed monotonicity of any sensible estimator, $\widehat{p}(x) < \widehat{p}(x')$ although x is considered more extreme from $\mathcal{H}_0$ than x' under $\succeq'$. Therefore, for a one-stage design, the only sensible definition of $\succeq$ is the natural ordering on the real numbers, i.e, $\succeq := \geq$.

The choice of estimator is also fairly canonical in the one-stage setting since the maximum likelihood estimator of the response probability, $\widehat{p}_{\text{MLE}}(x) := x/n$ is also unbiased. There are well-known correspondences between the decision to reject the null hypothesis ($X > c$), p values, and the unbiased maximum likelihood estimator, namely

$$\widehat{p}_{\text{MLE}}(x) > \widehat{p}_{\text{MLE}}(c) \quad \Leftrightarrow \quad X > c \quad \Leftrightarrow \quad \rho \leq \alpha \quad \Leftrightarrow \quad \text{reject } \mathcal{H}_0\,. \tag{5.4}$$

To quantify the uncertainty about the point estimate obtained from $\widehat{p}_{\text{MLE}}$ a two-sided confidence interval can be used. Here, a symmetric $1-2\alpha$ two-sided confidence interval is an interval-estimator

$$\left[\,\widehat{l}, \widehat{u}\,\right] : \{0, \dots, n\} \to \left\{\, (p, p') \in [\,0, 1\,] \,|\, p \leq p' \,\right\};\ \left[\,\widehat{l}, \widehat{u}\,\right](x) := \left[\,\widehat{l}(x), \widehat{u}(x)\,\right] \tag{5.5}$$

$$\Pr_p\left[\,\widehat{l}(X) > p\,\right] \leq \alpha \quad \wedge \quad \Pr_p\left[\,\widehat{u}(X) < p\,\right] \leq \alpha \qquad \forall p \in [\,0, 1\,] \tag{5.6}$$

where the last property encodes the defining coverage property of a symmetric confidence interval. Due to $X > c \Leftrightarrow \rho \leq \alpha$, a confidence interval can then be defined in terms of the p values alone. A canonical choice for $\left[\,\widehat{l}, \widehat{u}\,\right](\cdot)$ is given by the Clopper-Pearson confidence interval for binomial proportions (Clopper *et al.*, 1934), the pointwise solution of

$$\widehat{l}^{\,\text{CP}}(x) := \underset{\widehat{l}(x)}{\text{argmin}} : \quad \Pr_{\widehat{l}(x)}\left[X \geq x\right] > \alpha \tag{5.7}$$

$$\widehat{u}^{\,\text{CP}}(x) := \underset{\widehat{u}(x)}{\text{argmax}} : \quad \Pr_{\widehat{u}(x)}\left[X \leq x\right] > \alpha\,. \tag{5.8}$$

By definition, for any $p \in [0, 1]$,

$$\Pr_p\left[\,p < \widehat{l}^{\,\text{CP}}(X)\,\right] \leq \Pr_p\left[\{x = 0, \dots, n \,|\, \Pr_p\left[X \geq x\right] \leq \alpha\}\right] \leq \alpha \tag{5.9}$$

and similarly $\Pr_p\left[\,p > \widehat{u}^{\,\text{CP}}(X)\,\right] \leq \alpha$. Thus, the Clopper-Pearson interval is indeed exact as it guarantees a two-sided coverage of at least $1-2\alpha$ and a one-sided coverage of $1 - \alpha$ for both tails. Note that equations (5.7) and (5.8) are exclusively defined in

terms of one-sided p values since $\Pr_{\widehat{l}(x)}\left[X \geq x\right]$ is the p value for $\mathcal{H}_0 : p \leq \widehat{l}(x)$ (superiority test) and *vice versa* $\Pr_{\widehat{u}}\left[X \leq x\right]$ is the p value for $\mathcal{H}_0 : p \geq \widehat{u}(x)$ (inferiority test). Thus, it also holds true that

$$\rho(X) \leq \alpha \quad \Leftrightarrow \quad \widehat{l}^{\text{CP}}(X) > p_0 \,. \tag{5.10}$$

This corresponds to the common assumption that a confidence interval contains the boundary of the null hypothesis if and only if the p value is lower or equal to α which, in turn, is equivalent to rejecting the null hypothesis.

For a one-stage binomial test, frequentist inference is thus canonical and the different quantities (p value, estimator, and confidence interval) are compatible with each other and the underlying test decision.

5.2. Frequentist inference in two-stage designs

In two-stage designs, however, the carefully constructed frequentist inference framework easily leads to incompatibilities. To see this, it is important to note that both the definition of a p value and thus the corresponding Clopper-Pearson confidence interval crucially depend on the ordering $\succeq$ of the sample space. For a one-stage experiment binomial experiment this ordering is canonical and there is only one sensible choice, $\succeq = \geq$.

In a two-stage design, however, final outcomes are two-dimensional $(X_1, X_2) \in \mathbb{Z}^2$ and there is no canonical ordering on $\mathbb{Z}^2$ that would give rise to a unique definition of a p value. The analogous definition of a two-stage p value for a two-stage design is

$$\rho(x_1, x_2) := \Pr_{p_0}\left[\, (X_1, X_2) \succeq (x_1, x_2) \,\right], \tag{5.11}$$

only that the choice of the ordering '$\succeq$' on the outcome space $\mathbb{X}_{\mathcal{D}} \subset \mathbb{Z}^2$ of the design $\mathcal{D}$ is no longer unique.

The arbitrariness of p values for multi-stage designs is well known in the literature on group-sequential and adaptive designs (Jennison *et al.*, 2000, pp. 179). Different approaches for resolving this ambiguity have been proposed. The focus is primarily on the definition of a valid p value by heuristically justifying an ordering on the sample space (e.g. likelihood ratio ordering or score test ordering (Cook, 2002)). While most of these orderings are intuitively sensible, the mere fact that they are no longer equivalent for all two-stage designs (as it is the case for one-stage designs) implies that 'the' p value is not well-defined for multi-stage designs and depends on the choice of ordering. Jennison *et al.* (2000, p. 181) note that

> *'[a]lthough it is unfortunate that the definition of a P-value should depend on a choice of ordering [...], it should be stressed that different orderings yield very similar P-values for many outcomes.'*

It remains the fact, though, that the definition of the p value, which is often seen as *the* pivotal frequentist quantity summarising the results of a trial, is not well-defined for two-stage designs. An exhaustive discussion of all possible orderings or the situations

in which they lead to diverging *post-hoc* inference is beyond the scope of this thesis. A particularly interesting way of defining an ordering, however, is given by estimator-induced orderings (Jennison *et al.*, 2000; Cook, 2002; Kunzmann *et al.*, 2017a). Recall, that one of the many equivalent ways to justify the natural ordering on $\{0, \dots, n\}$ in the one-stage situation is to define

$$x \succeq_{\widehat{p}} x' : \quad \Leftrightarrow \quad \widehat{p}\,(x) \geq \widehat{p}\,(x') \,. \tag{5.12}$$

This definition can be seen as a way to pull back the natural ordering on $[\,0, 1\,] \subset \mathbb{R}$ to the sample space $\mathbb{X}_{\mathcal{D}}$ via an estimator $\widehat{p}$. The rationale for this ordering is that any observation with a larger associated point estimate is more extreme from $\mathcal{H}_0 : p \leq p_0$. Analogous to the one-stage case, any alternative ordering $\succeq' \neq \succeq_{\widehat{p}}$ necessarily implies the existence of a pair of observations $(x_1, x_2), (x_1', x_2')$ for which (x_1, x_2) is considered more extreme from the null hypothesis than (x_1', x_2') ,$(x_1, x_2) \succeq (x_1', x_2')$, yet the corresponding point estimates are ordered exactly in the opposite way, i.e. $\widehat{p}\,(x_1, x_2) < \widehat{p}\,(x_1', x_2')$. As a consequence, any ordering on $\mathbb{X}_{\mathcal{D}}$ that differs from $\succeq_{\widehat{p}}$ implies that the evidence against $\mathcal{H}_0$ as measured by the p value is incompatible, for at least one pair of observations, with the interpretation of a larger point-estimate corresponding to more evidence against $\mathcal{H}_0$. Since p value and point estimate are routinely reported together, compatibility of their interpretation in terms of evidence against $\mathcal{H}_0$ should be a desirable property.

A further complication in the choice of $\succeq$ arises when simultaneously taking into account the test decision of the underlying two-stage design. Any one-sided p value also induces a valid level-α test for the null hypothesis $\mathcal{H}_0 : p \leq p_0$ by rejecting whenever $\rho(x_1, x_2) \leq \alpha$. Depending on $\mathcal{D}$ and the choice of $\succeq$, the decision reached by the p value induced test need not be the same as the one implied by $\mathcal{D}$, i.e. there might be an outcome (x_1, x_2) with $\rho(x_1, x_2) \leq \alpha$ but $x_2 \leq c_2(x_1)$. Although, mathematically, this is not a problem *per se*, these situations tend to be a major practical nuisance when reporting the trial outcome as the equivalences given in equation (5.4) tend to be perceived as generally valid although they depend on the uniqueness of the sample space ordering. Since the construction of a confidence interval also depends on the ordering, similarly, its boundaries might overlap with the null hypothesis while the design rejects the null hypothesis or *vice versa*.

5.3. Unbiased estimation

To illustrate the issues raised above, a point estimator needs to be chosen. Since bias is often a major concern in the clinical trials community and all previously discussed Bayesian point estimators fail to completely eliminate it, it seems only natural to consider an unbiased estimator. It is well-known that the maximum likelihood estimator is biased in a two stage design (Bauer *et al.*, 2016). A completely unbiased estimator can be derived by applying the Rao-Blackwell theorem to improve the unbiased stage-one maximum likelihood estimator X_1/n_1. Jung *et al.* (2004) derived this unbiased Rao-Blackwell estimator in the group sequential case and the construction can directly be transferred to the slightly more involved case of generic two-stage designs (Kunzmann *et al.*, 2017a).

The principal idea of Rao-Backwellisation is to start with an unbiased estimator, X_1/n_1 and condition it on a sufficient statistic. The resulting estimator is then also unbiased and its variance is smaller or equal to the variance of the original estimator. Using Neyman's Factorization Lemma it is clear that $(n_2(X_1), X)$ is a sufficient statistic for the true response rate p since

$$\begin{aligned}
&\Pr_p[\, X_1 = x_1, X_2 = x_2 \,] && (5.13)\\
&= \Pr_p[\, X_2 = x_2 \,|\, X_1 = x_1 \,] \, \Pr_p[\, X_1 = x_1 \,] && (5.14)\\
&= \binom{n_2(x_1)}{x_2} p^{x_2}(1-p)^{n_2(x_1)-x_2} \binom{n_1}{x_1} p^{x_1}(1-p)^{n_1-x_1} && (5.15)\\
&= \underbrace{\binom{n_1}{x_1}\binom{n_2(x_1)}{x_2}}_{=:\, h(x_1,x_2)} \underbrace{p^x(1-p)^{n(x_1)-x}}_{=:\, g_p(n_2(x_1),x)} && (5.16)
\end{aligned}$$

where $h, g_p > 0$. Thus the Rao-Blackwellised estimator is given by

$$\widehat{p}_{\text{RB}} := \mathrm{E}_p \left[\, X_1/n_1 \,|\, (n_2(X_1), X) = (n_2, x)' \,\right] . \tag{5.17}$$

The probability mass function for the joint distribution of $(n_2(X_1), X)$ is

$$\begin{aligned}
&\Pr_p \left[\, (n_2(X_1), X) = (n_2, x) \,\right] && (5.18)\\
&= \sum_{\substack{x_1'=0 \\ n_2(x_1')=n_2}}^{n_1} \Pr_p \left[\, X_2 = x - x_1 \,|\, X_1 = x_1 \,\right] \Pr_p \left[\, X_1 = x_1 \,\right] && (5.19)\\
&= \sum_{\substack{x_1'=0 \\ n_2(x_1')=n_2}}^{n_1} \binom{n_1}{x_1} \binom{n_2(x_1)}{x - x_1} p^x(1-p)^{n(x_1)-x} . && (5.20)
\end{aligned}$$

Consequently,

$$\widehat{p}_{\text{RB}}(x_1, x_2) = \sum_{\substack{x_1'=0 \\ n_2(x_1')=n_2}}^{n_1} \frac{\frac{x_1'}{n_1}\binom{n_1}{x_1'}\binom{n_2(x_1')}{x-x_1'}\rho^x(1-\rho)^{n(x_1')-x}}{\sum_{\substack{x_1''=0 \\ n_2(x_1'')=n_2}}^{n_1}\binom{n_1}{x_1''}\binom{n_2(x_1'')}{x-x_1''}\rho^x(1-\rho)^{n(x_1'')-x}} \tag{5.21}$$

$$= \frac{\rho^x(1-\rho)^{n(x_1)-x}\sum_{\substack{x_1'=0 \\ n_2(x_1')=n_2}}^{n_1}\frac{x_1'}{n_1}\binom{n_1}{x_1'}\binom{n_2(x_1')}{x-x_1'}}{\rho^x(1-\rho)^{n(x_1)-x}\sum_{\substack{x_1=0 \\ n_2(x_1'')=n_2}}^{n_1}\binom{n_1}{x_1''}\binom{n_2(x_1'')}{x-x_1''}} \tag{5.22}$$

$$= \frac{\sum_{\substack{x_1'=0 \\ n_2(x_1')=n_2}}^{n_1}\binom{n_1-1}{x_1'-1}\binom{n_2(x_1')}{x-x_1'}}{\sum_{\substack{x_1''=0 \\ n_2(x_1'')=n_2}}^{n_1}\binom{n_1}{x_1''}\binom{n_2(x_1'')}{x-x_1''}}\,. \tag{5.23}$$

Here it is assumed that $n_1 > 0$ and the usual conventions on the binomial coefficient

$$\binom{n}{x} := 0 \quad \text{if} \quad n < 0 \,|\, x \leq 0 \,|\, x > n \tag{5.24}$$

apply. For $n_2 = n_2(x_1) = 0$ (early stopping), $\widehat{p}_{\text{RB}}(x_1, x_2)$ reduces to

$$\frac{\sum_{\substack{x_1'=0 \\ n_2(x_1')=0}}^{n_1}\binom{n_1-1}{x_1'-1}\binom{n_2(x_1')}{x-x_1'}}{\sum_{\substack{x_1''=0 \\ n_2(x_1'')=0}}^{n_1}\binom{n_1}{x_1''}\binom{n_2(x_1'')}{x-x_1''}} = \frac{\frac{x_1}{n_1}\sum_{\substack{x_1'=0 \\ n_2(x_1')=0}}^{n_1}\binom{n_1}{x_1'}}{\sum_{\substack{x_1''=0 \\ n_2(x_1'')=0}}^{n_1}\binom{n_1}{x_1''}} = \frac{x_1}{n_1} \tag{5.25}$$

i.e. to the stage-one maximum likelihood estimator. Note that this is also the case whenever the pre-image of n_2 under the designs sample size function $n_2(\cdot)$ is unique. To see this, let x_1^* be the unique pre-image of n_2 under $\mathcal{D}$, i.e. $n_2(x_1) = n_2 \Rightarrow x_1 = x_1^*$, then

$$\frac{\sum_{\substack{x_1'=0 \\ n_2(x_1')=n_2}}^{n_1}\binom{n_1-1}{x_1'-1}\binom{n_2(x_1')}{x-x_1'}}{\sum_{\substack{x_1''=0 \\ n_2(x_1'')=n_2}}^{n_1}\binom{n_1}{x_1''}\binom{n_2(x_1'')}{x-x_1''}} = \frac{\frac{x_1}{n_1}\binom{n_1}{x_1^*}\binom{n_2(x_1^*)}{x-x_1^*}}{\binom{n_1}{x_1^*}\binom{n_2(x_1^*)}{x-x_1^*}} = \frac{x_1}{n_1}\,. \tag{5.26}$$

For any two-stage design where $n_2(\cdot)$ is injective on the continuation region, this implies that $\widehat{p}_{\text{RB}}$ reduces to the stage-one maximum likelihood estimator and its estimates are consequently completely independent of any stage-two data. This observation has two consequences. Firstly, the variance of $\widehat{p}_{\text{RB}}$ is relatively large since it makes little or no use of any stage-two data. Secondly, consider the case of a non-trivial two-stage design where $n_2(\cdot)$ is injective on the continuation region. Here, non-trivial means that the test decision is not constant over all possible second stages. It is then clear that there must be outcomes $(x_1, x_2) \neq (x_1, x_2')$ where w.l.o.g. (x_1, x_2) leads to the rejection of the null hypothesis and (x_1, x_2') does not but $\widehat{p}_{\text{RB}}(x_1, x_2) = \widehat{p}_{\text{RB}}(x_1, x_2')$ since the stage-one outcomes are identical. For any such design, p values based on the ordering $\succeq_{\widehat{p}_{\text{RB}}}$ are thus necessarily incompatible with the design's test decision in the above discussed sense. Most generic optimal two-stage designs considered so far indeed have an injective stage-two sample size function on their respective continuation region. In fact, in cases where this is not the case, a simple group-sequential design would often be sufficient. This implies that unbiased inference and optimal response-adaptivity of the sample size function are more or less mutually exclusive.

The trade-offs of the different estimation techniques are compared in Section 11.1. Section 11.2 takes a closer look at the issue of design incompatible p values by means of a design proposed in Shan *et al.* (2016).

5.4. Compatible frequentist inference

The preceding sections raise the question whether a frequentist inferential framework for two-stage designs that satisfies the previously discussed compatibility properties between test decision, p value, point estimate, and confidence interval

$$\widehat{p}(X_1, X_2) > \widehat{p}(X_1, c_2(X_1)) \quad \Leftrightarrow \quad X_2 > c_2(X_1) \quad \Leftrightarrow \quad \rho(X_1, X_2) \leq \alpha \tag{5.27}$$

as well as the compatibility between the evidential interpretation of point estimate and p value

$$\widehat{p}(X_1, X_2) > \widehat{p}(X_1', X_2') \quad \Leftrightarrow \quad \rho(X_1, X_2) \leq \rho(X_1', X_2') \tag{5.28}$$

can exist. Note that such a framework is by no means necessary when test decision, p value, point estimate, and confidence interval are seen as unrelated entities. In practice, however, these compatibility properties are often presumed by the non-statistical readership of published trial results and ambiguous situation can be avoided by using a compatible framework.

At the core of the problem lies the definition of a suitable ordering for the sample space $\mathbb{X}_{\mathcal{D}}$. Any ordering implies a definition of the p value function and the corresponding confidence interval as outline in Section 5.2. To achieve compatibility between point estimates and p values, the ordering must be implied by the chosen point estimator. That is, if the point estimator is chosen such that its induced p values are compatible with the underlying design's test decision, all of the above properties hold. To solve the problem of an overall compatible framework for frequentist inference in a two-stage design, one thus needs to explore the relation between $\widehat{p}$ and the decision criterion $X_2 > c_2(X_1)$ of the level-α design $\mathcal{D}$ for testing $\mathcal{H}_0 : p \leq p_0$.

Let $\mathbb{X}_{\mathcal{D}}^{+} \subset \mathbb{X}_{\mathcal{D}}$ be the rejection region of $\mathcal{D}$ and let $\mathbb{X}_{\mathcal{D}}^{-} = \mathbb{X}_{\mathcal{D}} - \mathbb{X}_{\mathcal{D}}^{+}$ be the complementary region of non-rejection. Any estimator $\widehat{p}$ that induces a compatible inferential framework must lead to a perfect separation of the disjoint subspaces $\mathbb{X}_{\mathcal{D}}^{+}$ and $\mathbb{X}_{\mathcal{D}}^{-}$ in the sense that

$$(x_1, x_2) \in \mathbb{X}_{\mathcal{D}}^{+} \quad \Rightarrow \quad \left\{ (x_1', x_2') \in \mathbb{X}_{\mathcal{D}} \mid \widehat{p}\,(x_1', x_2') \geq \widehat{p}\,(x_1, x_2) \right\} \subset \mathbb{X}_{\mathcal{D}}^{+} \,. \tag{5.29}$$

This property is both sufficient and necessary for compatibility of the test induced p values and the design $\mathcal{D}$. Since $\mathcal{D}$ was assumed to define a level-α test, compatibility of $\widehat{p}$ and $\mathcal{D}$ trivially implies the weaker

$$(x_1, x_2) \in \mathbb{X}_{\mathcal{D}}^{+} \quad \Rightarrow \quad \Pr_{p_0}\left[\widehat{p}\,(X_1, X_2) \geq \widehat{p}\,(x_1, x_2) \right] \leq \alpha \,. \tag{5.30}$$

This means that any test induced by a compatible estimator is also a level α test for $\mathcal{H}_0 : p \leq p_0$.

An obvious idea to address the compatibility issue is to impose additional constraints on the optimal design that guarantee compatibility with the estimator of choice. To achieve test compatibility, one would need to make sure that

$$\begin{aligned} x_2 > c_2(x_1) \wedge x_2' \leq c_2(x_1') \Rightarrow \widehat{p}\,(x_1, x_2) > \widehat{p}\,(x_1', x_2') \\ \forall\, (x_1, x_2), (x_1', x_2') \in \mathbb{X}_{\mathcal{D}} \,. \end{aligned} \tag{5.31}$$

This set of pairwise implications leads to perfect separation of the rejection region by $\widehat{p}_{\text{MLE}}$ and thus to compatibility of test decision with $\succeq_{\widehat{p}_{\text{MLE}}}$. In practice, implementing these pairwise consistency checks during optimisation of the design is infeasible due to the large number of possible pairs for all possible design configurations and reverse dependency of some estimators on the design itself (e.g. the unbiased estimator). Instead, compatibility can only be checked *post hoc* in practice.

However, the concept of compatibility between design $\mathcal{D}$ and estimator $\widehat{p}$ (and thus the induced p value function and confidence interval) is symmetrical. Just as a design can theoretically be forced to be compatible with any given estimator, a compatible estimator can also be constructed for a given design. In the case of a binary two-stage design, an estimator is a function mapping from a finite space $\mathbb{X}_{\mathcal{D}}$ to the unit interval. Conveniently, $\widehat{p}$ can thus be understood as a finite set of real numbers

$$\widehat{p} \;=\; \widehat{p}_{\,\mathcal{D}} \;=\; \left\{ \widehat{p}\,(x_1, x_2) \in [\,0, 1\,] \mid (x_1, x_2) \in \mathbb{X}_{\mathcal{D}} \right\} \tag{5.32}$$

indexed by the respective observation (x_1, x_2). Since there are infinitely many such functions, an objective criterion must be chosen to define the $|\,\widehat{p}\,| \;=\; |\,\mathbb{X}_{\mathcal{D}}|$ estimates. For instance, the maximum likelihood estimator maximises the binomial likelihood point-wise leading to $\widehat{p}_{\text{MLE}}(x_1, x_2) = (x_1 + x_2)/n(x_1)$. Constraint (5.31), however, is global on the set $\widehat{p}$ and thus requires a global objective criterion to balance between local deviation from the likelihood maximiser and global constraints violations. To guarantee compatibility, a local deterioration of the likelihood fit for individual (x_1, x_2) must be tolerated to allow the estimator to comply with the compatibility constraints. One possible approach is to minimise the maximal local likelihood

difference between the compatible estimator and the (locally optimal) standard maximum likelihood estimator. Let the compatible maximum likelihood estimator $\widehat{p}_{\text{CMLE}}$ be the solution of

$$\underset{\widehat{p}(x_1,x_2)\in\widehat{p}}{\operatorname{argmin}} : \sup_{(x_1,x_2)\in\mathbb{X}_{\mathcal{D}}} \left(\ell\big(x_1,x_2,\widehat{p}\,(x_1,x_2)\big) - \ell\big(x_1,x_2,\widehat{p}_{\text{MLE}}(x_1,x_2)\big) \right)^2 \tag{5.33}$$

$$\text{subject to} : (x_1,x_2) \in \mathbb{X}_{\mathcal{D}}^{+}$$

$$\Rightarrow \left\{ (x_1',x_2') \in \mathbb{X}_{\mathcal{D}} \mid \widehat{p}\,(x_1',x_2') \geq \widehat{p}\,(x_1,x_2) \right\} \subset \mathbb{X}_{\mathcal{D}}^{+} \tag{5.34}$$

$$\forall\, (x_1,x_2) \in \mathbb{X}_{\mathcal{D}}^{+} : n_2(x_1) = n_2(x_1') \;\wedge\; x_1 = x_1'$$

$$\Rightarrow \widehat{p}\,(x_1,x_2) > \widehat{p}\,(x_1',x_2') \tag{5.35}$$

$$\forall\, (x_1,x_2) \in \mathbb{X}_{\mathcal{D}}^{+} : x_1 = x_1' \;\wedge\; x_2 > x_2'$$

$$\Rightarrow \widehat{p}\,(x_1,x_2) > \widehat{p}\,(x_1',x_2') \tag{5.36}$$

where $\ell(x_1,x_1,p)$ is the likelihood of observation (x_1,x_2) under response probability p. In cases where $\widehat{p}_{\text{MLE}}$ is already compatible, the constraints are non-binding and the solution reduces to the standard maximum likelihood estimator. In any other case, the resulting estimator will be a distorted maximum likelihood estimator where the degree of distortion depends on the severity of the constraint violations of $\widehat{p}_{\text{MLE}}$. Since any violation of the compatibility constraint may lead to global distortions of $\widehat{p}_{\text{MLE}}$, constraints (5.35) and (5.36) ensure that the new ordering still satisfies minimal plausible constraints (analogous to the stage-wise ordering (Jennison *et al.*, 2000)).

Note that the proposed method is more generally applicable. For instance, in (Kunzmann *et al.*, 2017a), an optimal compatible estimator minimising a weighted mean-square criterion was proposed. The weight function can be interpreted as a prior φ over p. The solution is thus a potentially distorted version of the posterior mean estimator discussed in Chapter 4, since the posterior mean minimises the quadratic Bayes risk. However, a modification of $\widehat{p}_{\text{MLE}}$ seems more appropriate in a frequentist setting to avoid the necessity of specifying a prior density φ. The only subjective choice in the proposed approach to derive $\widehat{p}_{\text{CMLE}}$ is the exact specification of the objective (5.33). The local deviation of $\widehat{p}_{\text{CMLE}}$ at (x_1,x_2) from the maximiser $\widehat{p}_{\text{MLE}}(x_1,x_2)$ could also be measured differently. For instance, instead of the squared difference in the respective likelihoods, the squared difference of the estimates themselves, $(\,\widehat{p}_{\text{CMLE}}(x_1,x_2) - \widehat{p}_{\text{MLE}}(x_1,x_2)\,)^2$, or the relative likelihood difference could be used. The advantage of the above proposed objective is that deviations from the MLE are naturally weighted by the likelihood of the corresponding observation. Numerically, problems of the class considered here are challenging. The dimension $|\,\mathbb{X}_{\mathcal{D}}|$ of the optimisation is potentially high dimensional (roughly 50 to 200 for the designs considered so far). Also, the objective function is non-linear, non-smooth (maximum!), and the number of constraints is considerable. Yet, the problem only involves continuous variables and can thus be solved using gradient based local optimisation methods. A natural starting point for the optimisation is given with $\widehat{p}_{\text{MLE}}$. Incompatibility with the MLE occurs rarely when considering optimal design obtained from minimising expected

sample size. An example is given in Section 11.3 and the corresponding compatible maximum likelihood estimator is compared with the vanilla MLE.

6. Unplanned Design Adaptations

6.1. Rationale for unplanned design adaptations

In the preceding chapters, it was implicitly assumed that the planning assumptions encoded in the prior φ remain valid throughout the conduct of a trial. In practice, however, planning assumptions can change. Even small phase II trials in early oncology may run for several years. In a fast-paced research environment, emerging new evidence or a reassessment of existing evidence may thus invalidate the original planning assumptions of a trial. Since the optimality of a design generally depends on the underlying planning prior, an optimal design can thus become obsolete during the course of its conduct.

It is important to note that no trial-internal data event could ever trigger the need for a design adaptation when an optimal design is employed. This is a direct consequence of the fact that *all possible* stage-one outcomes are considered during optimal planning. For instance, instead of recalculating the sample size in the event of undershooting a certain target value (see Section 1.4), a constraint on the minimal conditional power of the optimal design should be incorporated in the optimisation problem (see Section 2.3.1). This allows the optimisation process to avoid situations with low conditional power in an optimal way. By definition, an optimal design incorporating such additional constraints must thus always be superior to a *post hoc* adjustment of a non-optimal design. Formally, this inefficiency of unplanned adaptation is related to the fact that they need to invoke some form of the conditional error principle (see Section 1.4) to maintain strict type one error rate control. This imposes additional constraint on the (conditional) type one error rate of the recalculated design. Pre-specified optimal two-stage designs, however, are only restricted by the unconditional error rate constraints and may thus be more effective since the feasible space of the optimisation problem is larger. Thus, designs obtained from the binding application of a recalculation rule are necessarily less effective than a completely pre-specified design directly optimising the chosen objective criterion. Unplanned adaptations for optimal two-stage designs are therefore only justified if a deviation from the pre-specified sampling scheme is required or new *trial-external* information becomes available. If this is indeed the case, the original planning prior φ needs to be updated to a new 'prior' φ' that reflects the change of trial external information. In this context, the temporal connotation of the term 'prior' is somewhat misleading. A more appropriate interpretation is that the prior serves as a quantification of any trial-external information - which might naturally change during the course of the trial.

Since the only reasons to adapt an optimally planned design are trial-external, the change of design could be assumed to be stochastically independent of any trial-

internal information. Yet, decision-makers are rarely completely agnostic of already accrued response data in single-arm trials since blinding is impossible. The assumption that potential adaptation decisions are data-independent and would not affect error rate control is therefore hard to justify in practice. For instance, the mere decision whether to consider new information as relevant to the particular phase II trial at hand and update the prior accordingly introduces additional degrees of freedom to decision makers. Instead, a framework for unplanned adaptations in single-arm trials needs to assume that the entire trial-internal data is available to decision makers (Englert *et al.*, 2015).

Methods for unplanned adaptations under strict type one error rate control were already discussed in Section 1.4. The distinction between the methodology for unplanned design changes ('sample size recalculation') and the response adaptivity of the sample size functions developed in Chapters 2 and 3 is crucial. The latter can, in effect, be understood as a form of randomised test where the randomisation is based on assumptions about the distribution of the unobserved interim outcome and collapses to a deterministic decision given the interim results. In terms of practical acceptance, this trial-internal randomisation is much easier to justify than one that is based on a trial-external biased coin toss. The analogy with randomised testing procedures makes it clear that generic two-stage designs with response adaptive sample size function are still completely pre-specified testing procedures. As such, the operating characteristics of these designs are well understood and the design can be tailored to fulfil error rate constraints in an optimal way which is governed by an objective function. The theoretical justification for an *unplanned* adaptation, however, is much more involved (see Section 1.4). The main tool for implementing such changes in an ongoing trial is the conditional error principle, which specifies a sufficient condition for satisfying a strict error rate constraint under mild technical assumptions (Brannath *et al.*, 2012). This powerful methodology allows a *'frightening multitude of flexibility'* (Bauer *et al.*, 2016) of design adaptations. A multitude of sensible heuristics like the one introduced in Section 1.4 were put forward (Proschan *et al.*, 1995; Bauer *et al.*, 2016). Still, the *ideal* choice of the adaptation criterion for an unplanned adjustment of a pre-planned design remains an open problem. In this Chapter, a suggestion is put forward as to how a given optimal design can be adjusted in an unplanned manner while respecting the properties of the original optimisation criterion.

6.2. Notation for unplanned design adaptations

A thorough discussion of unplanned design adaptations requires a slightly more flexible data model. Let, to that end, $(R_i)_{i\geq 1}$ be a time-discrete Bernoulli process with response probability p modelling the (possibly) infinite number of responses (a response of the i-th individual is encoded as $R_i = 1$) observed within a trial. Further, let $\mathcal{D}^* = (\, n_1^*, n_2^*(\cdot),^* c_2(\cdot)\,)$ be the original design optimised for a score s under constraints on the maximal type one error rate (at p_0) and expected power given a planning prior φ and a minimal clinically relevant response probability of $p_{\text{MCR}} \geq p_0$, i.e.,

$\mathcal{D}^*$ is the solution of

$$\underset{n_1,\, n_2(\cdot),\, c_2(\cdot)}{\operatorname{argmin}} \; : \; \mathrm{E}_{\varphi}[\, s(\mathcal{D}, X_1, X_2, p)\,] \tag{6.1}$$

$$\text{subject to :} \quad \Pr_{p_0}[\, X_2 > c_2(X_1)\,] \leq \alpha \tag{6.2}$$

$$\Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1) \,|\, p \geq p_{\text{MCR}}\,] \geq 1 - \beta\,. \tag{6.3}$$

The definition of the stage-wise test statistics X_1 and X_2 in terms of the data-generating process $(R_i)_{i\geq 1}$ can be recovered as $X_1 = X_{(0,n_1]}$ and $X_2 = X_{(n_1,n_2(X_1)]}$ where

$$X_{(a,b]} := \sum_{i=a+1}^{b} R_i\,. \tag{6.4}$$

Let $\tau = n(X_1)$ be the (random) stop time for the process (R_i) under $\mathcal{D}^*$. If the stop time τ is observed, the trial ended according to the pre-specified design $\mathcal{D}^*$. If however, the data-generating process (i.e., patient recruitment) is stopped earlier for some $\tau' < \tau$, an unplanned interim analysis can be conducted based on the data accrued so far, $(R_i)_{i=0,\ldots,\tau'} = (r_1, r_2, \ldots, r_{\tau'})$. Further assume that the prior might have changed from φ to φ' to reflect new trial-external information[1]. In this situation, the issue is to determine an design modification conditional on the already observed data to reflect the change in external information while maintaining strict type one error rate control.

6.3. Optimal unplanned adaptations in stage two

The simplest case of an unplanned adaptation occurs if the first stage of the original design was already completed, i.e., $\tau' \geq n_1$. The slightly more involved case of an adaptation in stage-one is discussed in the next Section. Then, $n_2(x_1)$ is known and the adaptation problem reduces to finding new $n_2'(x_1)$ and $c_2'(x_1)$.

A natural choice for the objective function of the adaptation problem is the conditional expected score given the data observed up to and including time point τ' under the new prior φ', i.e.

$$\mathrm{E}_{\varphi'}[\, s(\mathcal{D}', X_1', X_2', p) \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, r_2, \ldots, r_{\tau'})\,]\,, \tag{6.5}$$

where $X_1' := X_{(0,n_1]}$ and $X_2' := X_{(n_1,n_2'(X_1)]}$ are the stage-wise test statistics under the modified design . Here, $X_1' = X_1$ holds since stage one was already completed. Expression (6.5) only depends on the modified design via $n_2'(x_1)$ and $c_2'(x_1)$ since x_1 is observed. I.e., the remainder of the modified design $\mathcal{D}'$ remains unspecified. Similarly, maximal type one error rate and (expected) power can be evaluated conditional on $(R_i)_{i=0,\ldots,\tau'} = (r_1, r_2, \ldots, r_{\tau'})$. The conditional error principle may then be applied

[1] All subsequent arguments remain valid if $\varphi' = \varphi$ (no change in prior) but it should be stressed that there is generally no need to perform an adaptation when the trial external information does not change.

to maintain strict maximal type one error rate control by restricting the conditional error of the modified design to

$$\begin{aligned} &\Pr_{p_0}[\, X_2' > c_2'(X_1') \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, r_2, \ldots, r_{\tau'})\,] \\ \leq &\Pr_{p_0}[\, X_2 > c_2(X_1) \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, r_2, \ldots, r_{\tau'})\,] \,. \end{aligned} \tag{6.6}$$

Since this constraint ensures that the conditional maximal type one error rate of the new design is less than the old design's it guarantees overall maximal type one error rate control at level α (Müller *et al.*, 2004; Brannath *et al.*, 2012).

For conditional power, most authors propose to impose a hard constraint using the original threshold of $1-\beta$ from the planning stage (Bauer *et al.*, 2016) (see also Section 1.4). Proschan *et al.* (1995) kept the choice of the conditional power threshold open, i.e., they allow a different threshold $1-\beta'$ for conditional power at the unplanned interim analysis.

$$1 - \Pr_{\varphi'(\cdot)}[\, X_2' > c_2'(X_1') \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, r_2, \ldots, r_{\tau'})\,] \leq 1-\beta' \,. \tag{6.7}$$

There are good reasons to modify the threshold on conditional expected power during the interim analysis. While it is certainly desirable to have a high expected probability to reject the null hypothesis given an relevant effect in stage two, the naïve application of the original threshold of $1-\beta$ may lead to excessively large sample sizes whenever the data observed before the interim analysis are supporting the null hypothesis. In Section 1.4 this was countered by imposing a hard constraint on the maximal allowable recalculated sample size. Consequently, for very low x_1, the recalculated design does not achieve a conditional power of $1-\beta$.

A more natural way to obtain a situation specific value of β' is given via the conditional error principle. Its application in the literature is restricted to controlling the unconditional type one error rate under recalculation. However, there is no reason why it should not be employed in the same way to (expected) type two error and impose the following conditional expected power constraint during recalculation

$$\begin{aligned} &\Pr_{\varphi'(\cdot)}[\, X_2' > c_2'(X_1') \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, r_2, \ldots, r_{\tau'})\,] \\ \geq\ &\Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1) \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, r_2, \ldots, r_{\tau'})\,] \,. \end{aligned} \tag{6.8}$$

This is merely an application of the conditional error principle to the (expected) type two error rate and sets $1-\beta' = \Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1) \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, r_2, \ldots, r_{\tau'})\,]$. This means that the conditional expected power of the recalculated design (under the new prior) is to be chosen at least as large as the conditional expected power of the old design (under the old prior). Since interim data that does not provide evidence against the null hypothesis would also lead to a low conditional expected power under the original design and prior, the threshold for the recalculation is naturally lowered as well. *Vice versa*, if the original design had a higher conditional expected power for the given interim data than $1-\beta$, the threshold for the recalculated design is also increased. A subtle detail of inequality (6.8) is the fact that the conditional power of the original design, $\mathcal{D}$, is evaluated under the original planning prior φ. Otherwise, if $\mathcal{D}$ was severely underpowered under φ', so

would the modified second stage. Both conditional error rate constraints (6.6) and (6.8) together imply that $\tau' < \tau$. This is due to the fact that, for any $p \in [0, 1]$, $\Pr_p[\, X_2 > c_2(X_1) \,|\, (R_i)_{i=0,\dots,\tau} = (r_1, r_2, \dots, r_\tau)\,] = \mathbf{1}_{x_2 > c_2(x_1)}$. That is, after observing $\tau = n_2(X_1)$ individuals ($\mathcal{D}$ completed) the decision under $\mathcal{D}$ is deterministic and the conditional power as function of p is either constantly 0 ($\mathcal{D}$ fails to reject) or constantly 1 ($\mathcal{D}$ rejects). To fulfil both conditional error rate constraints, $\mathcal{D}'$ must therefore lead to the same decision based on $(R_i)_{i=1,\dots,\tau}$ and the collection of any further data cannot change the outcome.

The problem of finding the modified stage-two sample size $n_2'(x_1)$ and the modified stage-two critical value $c_2'(x_1)$ using the proposed approach can thus be expressed as optimisation problem conditional on the data observed before the interim adaptation. The problem

$$\underset{n_2'(x_1),\, c_2'(x_1)}{\operatorname{argmin}} \;:\; \mathrm{E}_{\varphi'}[\, s(\mathcal{D}', X_1', X_2', p) \,|\, (R_i)_{i=0,\dots,\tau'} = (r_1, \dots, r_{\tau'})\,] \tag{6.9}$$

$$\begin{aligned} \text{subject to :}\;\; & \Pr_{p_0}[\, X_2' > c_2'(X_1') \,|\, (R_i)_{i=1,\dots,\tau'}\,] \\ & \leq \Pr_{p_0}[\, X_2 > c_2(X_1) \,|\, (R_i)_{i=0,\dots,\tau'} = (r_1, \dots, r_{\tau'})\,] \qquad (6.10) \\ & \Pr_{\varphi'(\cdot)}[\, X_2' > c_2'(X_1') \,|\, (R_i)_{i=0,\dots,\tau'} = (r_1, \dots, r_{\tau'})\,] \\ & \geq \Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1) \,|\, (R_i)_{i=0,\dots,\tau'} = (r_1, \dots, r_{\tau'})\,] \qquad (6.11) \end{aligned}$$

only has two free integer-valued variables, $n_2'(x_1)$ and $c_2'(x_1)$, and can easily be solved by an exhaustive search over both.

The key to computing both the conditional error rates and the conditional objective is the conditional distribution of $X_2 \,|\, (R_i)_{i=0,\dots,\tau'} = (r_1, \dots, r_{\tau'})$. Since $\tau' \geq n_1$, $X_2 \,|\, (R_i)_{i=0,\dots,\tau'} = (r_1, \dots, r_{\tau'}) = X_2 \,|\, X_1 = x_1,\, X_{(n_1,\tau']} = x_{(n_1,\tau']}$ and

$$\begin{aligned} & \Pr_p[\, X_2 = x_2 \,|\, X_1 = x_1,\, X_{(n_1,\tau']} = x_{(n_1,\tau']}\,] \\ = \; & \Pr_p[\, X_{(\tau',n_2(x_1)]} = x_2 - x_{(n_1,\tau']}\,] \qquad (6.12) \end{aligned}$$

Here, $X_{(\tau',n_2(x_1)]}$ is binomially distributed with size-parameter $n_2(x_1) - \tau' - n_1$ and probability p. This allows computing the conditional error rates and the conditional expected objective for both the original and the adapted design in the same way. E.g., the conditional type one error rate of the original design is

$$\begin{aligned} & \Pr_{p_0}[\, X_2 > c_2(X_1) \,|\, (R_i)_{i=0,\dots,\tau'} = (r_1, \dots, r_{\tau'})\,] \\ = \; & 1 - \Pr_{p_0}[\, X_{(\tau',n_2(x_1)]} \leq c_2(x_1) - x_{(n_1,\tau']}\,]\,. \qquad (6.13) \end{aligned}$$

The process of conducting a sample size adaptation in stage-two is illustrated in Section 12.2.

6.4. Optimal unplanned adaptations in stage one

The adaptation problem is slightly more involved when $\tau' < n_1$, i.e., when the adaptation occurs before proceeding to stage two, or if it is decided to prolong the first

stage beyond the original n_1. In both cases, the final sample size and the stage-two critical value remain to be determined in a future interim analysis. The result of an unplanned design adaptation must then be a new *partial* design

$$\mathcal{D}'|_{x_1' \geq x_{(0,\tau')}} = \left(n_1', n_2'|_{x_1' \geq x_{(0,\tau')}}(\cdot), c_2'|_{x_1' \geq x_{(0,\tau')}}(\cdot) \right) \tag{6.14}$$

where the co-domain of the functions n_2' and c_2' is restricted to the still observable $x_1' \in \{x_{(0,\tau')}, \ldots, n_1'\}$. Furthermore $n_1' \geq \tau'$ must hold to ensure that the time point of the new interim analysis lies in the future. The problem is thus structurally similar to the unconditional optimisation problem (see Section 2.3) only that the co-domain of the functions c_2' and n_2' is restricted to numbers of responses that can still be observed within the modified design. With the same arguments as before, this results in the following optimisation problem

$$\begin{aligned}
\underset{\substack{n_1', \\ n_2'|_{x_1' \geq x_{(0,\tau')}}(\cdot), \\ c_2'|_{x_1' \geq x_{(0,\tau')}}(\cdot)}}{\operatorname{argmin}} \quad & : \ \mathrm{E}_{\varphi'(\cdot)}[\, s(\mathcal{D}', X_1', X_2', p) \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, \ldots, r_{\tau'}) \,] && (6.15) \\
\text{subject to} : \ & \Pr_{p_0}[\, X_2' > c_2'(X_1') \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, \ldots, r_{\tau'}) \,] \\
& \qquad \leq \ \Pr_{p_0}[\, X_2 > c_2(X_1) \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, \ldots, r_{\tau'}) \,] && (6.16) \\
& \Pr_{\varphi'(\cdot)}[\, X_2' > c_2'(X_1') \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, \ldots, r_{\tau'}) \,] \\
& \qquad \geq \ \Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1) \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, \ldots, r_{\tau'}) \,] \,. && (6.17)
\end{aligned}$$

The same solution strategy via integer linear programming as outlined in Section 2.3 can be applied.

Further considerations might be incorporated in the adaptation procedure. In some instances it might be desirable to exclude the possibility of a further interim analysis from the adapted design $\mathcal{D}'$. This might, be the case when the unplanned adaptation takes place shortly before the initially planned interim analysis. In this case, n_1' has to be restricted to the observed number of outcomes during the unplanned interim analysis ($n_1' = \tau'$) to ensure that the interim analysis of the adapted design coincides with the time point of adaptation. The recalculated partial design can then be directly applied after the recalculation to determine the stage-two sample size and critical value.

Note that under $\tau' < n_1$ no data in the second stage is observed and

$$\begin{aligned}
& \Pr_p[\, X_2 = x_2, X_1 = x_1 \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, \ldots, r_{\tau'}) \,] \\
= \ & \Pr_p[\, X_2 = x_2 \,|\, X_1 = x_1 \,] \, \Pr_p[\, X_1 = x_1 \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, \ldots, r_{\tau'}) \,] && (6.18) \\
= \ & \Pr_p[\, X_2 = x_2 \,|\, X_1 = x_1 \,] \, \Pr_p[\, X_{(\tau', n_1]} = x_1 - x_{(0,\tau']} \,] \,, && (6.19)
\end{aligned}$$

where $X_{(\tau', n_1]}$ is also binomially distributed with size parameter $n_1 - \tau'$ and probability p. Therefore, the conditional type one error rate under the old design is given

by

$$\begin{aligned}
&\Pr_{p_0}[\, X_2 > c_2(X_1) \,|\, (R_i)_{i=0,\ldots,\tau'} = (r_1, \ldots, r_{\tau'}) \,] \\
&= \sum_{x_1 = x_{(0,\tau']}}^{n_1} \Pr_{p_0}[\, X_{(\tau',n_1]} = x_1 - x_{(0,\tau']} \,] \left(1 - \Pr_{p_0}[\, X_2 \leq c_2(x_1) \,] \right) \qquad (6.20)
\end{aligned}$$

and the expected score and expected power can be evaluated in an analogue way. Equation (6.20) shows that the conditional error of a design for a recalculation in stage one is simply the weighted sum of the conditional errors at the initially planned interim time point weighted with their respective chance of still being realised.

A worked example of this methodology is presented in Section 12.1.

6.5. Previous work

Englert and Kieser already considered unplanned design adaptations based on the representation of a design as conditional error function (Englert *et al.*, 2015). They addressed the slightly different problem of obtaining a new valid conditional error function (see Section 1.4 in the case of over- or underrunning the pre-planned stage one sample size). Underrunning here refers to a situation where the interim analysis is to be conducted at an earlier point in time than originally anticipated and overrunning to the situation where the interim analysis is delayed.

In contrast to just obtaining a conditional error function (i.e. a critical value for the final test decision) the above introduced framework uses the original optimisation criterion to produce a conditionally optimal stage-two sample size (adaptation in stage two) or a partial stage-two sample size function (adaptation in stage one). By linking the conditional recalculation back to the original planning problem it gives a natural answer to the question raised by Bauer *et al.* (2016) as to how the flexible recalculation methodology can be used in an *optimal* way. The proposed method can also be used to address mere over- or underrunning in stage two by simply keeping the original planning prior φ if no new trial-external information is available.

6.6. Inference after unplanned design adaptations

Frequentist inference generally depends on the entire design. For instance, the definition of a p value crucially depends on unobserved potential outcomes and the likelihood of these outcomes under the null hypothesis (see Section 5.2). Similarly, a mean-unbiased point estimator requires knowledge of the entire design up-front. Since no overall design is available after an unplanned adaptation but merely a partial one, frequentist inference based on p values and confidence intervals is no longer straight forward. The problem of obtaining valid p values after an adaptation could be addressed by pre-specifying a p value combination function with fixed weights during the design stage of the trial (see Section 1.4 for details on the combination function method). This allows the computation of valid p values under arbitrary design adaptation at the cost of introducing inefficiencies due to the pre-specified weights of

the combination function. A drawback of this approach is that compatibility of the p values (and thus the corresponding confidence intervals) with the final test decision can no longer be guaranteed.

Another possibility is to consider the above proposed unplanned adaptation rule as binding and to condition on the new prior φ' and the time point of the unplanned adaptation τ'. One may then compute the adapted design for all possible numbers of observed responses at the unplanned adaptation time τ'. This results in a well-defined sampling space and p values conditional on the adaptation rule, φ', and τ' can be computed based on, e.g., the MLE ordering. Since both τ' and φ' as well as the adaptation rule itself might be chosen data-dependently in practice, this approach is unlikely to be unequivocally acceptable though.

Bayesian inference on the other hand is independent of the sampling scheme. The posterior distribution of the response probability only depends on the final sample size and the overall number of responses observed over the course of the entire trial. Any quantity derived from the posterior distribution (point estimate, credible interval) therefore remain valid after unplanned adaptations.

7. Extension to the Continuous Case

The focus of this thesis is the rigorous treatment of the case of single-arm trials in oncology with binary endpoint. The general principles outlined in the previous chapters can, however, be applied to trials with other endpoints as well. This might be of interest when the primary endpoint of a trial is not tumour response but a continuous measure like differences in tumour volume. In the following, an outline of the approach for deriving optimal two-stage design for (approximately) normally distributed test statistics is presented. This chapter is a brief introduction to the work previously published in (Pilz *et al.*, 2019).

The general principle can be illustrated by means of a one-sided test for a mean difference with known variance. For instance, one could be interested in assessing whether the differences in tumour volume between baseline and a pre-specified follow-up time point differ between an intervention and a control group. Without loss of generality, assume that the endpoint of interest (within-patient tumour volume difference) has unit variance in both the intervention and the control group. The stage-wise test statistics X_1 and X_2 are then given by the standardised observed mean difference between intervention and control groups within each stage. Further assume that the underlying data generating mechanism is such that the central limit theorem can be invoked and that $X_1 \sim \mathcal{N}(\sqrt{n_1}\,\theta/\sqrt{2}, 1)$ and the stage-two test statistic $X_2 \,|\, n_2(x_1) > 0, X_1 = x_1 \sim \mathcal{N}(\sqrt{n_2(x_1)}\,\theta/\sqrt{2}, 1)$ where θ is the unknown standardised treatment difference. Let positive values of θ indicate superiority of the intervention group over the control group (i.e., one considers the difference 'control' - 'intervention'). Here, n_1 and $n_2(\cdot)$ correspond to the per-group sample sizes. Let c_1^f be the boundary for early futility stopping, c_1^e be the boundary for early efficacy stopping, and let $c_2(\cdot)$ be the stage-two critical value function such that the test rejects the null hypothesis if and only if $X_2 > c_2(X_1)$. In analogy to the binary case, it is assumed that $c_2(x_1) = \infty$ if $x_1 < c_1^f$ and $c_2(x_1) = -\infty$ if $x_1 > c_1^e$. To test $\mathcal{H}_0 : \theta \leq 0$ in a two-stage design, one then needs to derive optimal $n_1, n_2(\cdot), c_1^f, c_1^e$, and $c_2(\cdot)$.

For the sake of simplicity, further assume that the power at a point alternative $\theta_1 > \theta_0$ is restricted to a minimum of $1-\beta$, the maximal type one error rate is α, and that the objective criterion is expected sample size under θ_1. In analogy to Chapter 2 the optimisation problem is then

$$\underset{n_1, c_1^f, c_1^e, n_2(\cdot), c_2(\cdot)}{\operatorname{argmin}} : \quad \int_{-\infty}^{\infty} \phi(x_1 - \sqrt{n_1}/\sqrt{2}\,\theta_1)\, n(x_1)\, \mathrm{d}z_1 \tag{7.1}$$

$$\text{subject to :} \quad \Pr_{\theta_0}[\, X_2 > c_2(X_1)\,] \leq \alpha \tag{7.2}$$

$$\Pr_{\theta_1}[\, X_2 > c_2(X_1)\,] \geq 1-\beta\,. \tag{7.3}$$

This problem is similar to the one outlined in Section 2.3, only that both n_2 and c_2 now map from $\mathbb{R}$ to the natural numbers (n_2) and real values (c_2). Strictly speaking, the problem is still a mixed integer one, since n_1 and n_2 are integer-valued. For larger sample sizes, however, the approximation error incurred by assuming that both are also real-valued and then rounding to the nearest integer is small (Kunzmann *et al.*, 2020c). This approach was also used by Banerjee *et al.* (2006); Jennison *et al.* (2015).

Since n_2 and c_2 are functions, the problem is a variational one. It can either be addressed via the indirect Euler-Lagrange methodology by introducing Lagrange multipliers for the constraints before solving the corresponding Euler-Lagrange equation for each x_1 (Pilz *et al.*, 2019) or by parameterising the functions n_2 and c_2 on $[c_1^f, c_1^e]$ over a spline basis and solving for the optimal parameters directly (Kunzmann *et al.*, 2020c). In either case, the resulting problems are smooth in their respective parameters and standard optimisation methods can be used to obtain (numerical) solutions.

The same methodology can also be extended to asymptotically normally distributed test statistics to cover a wider range of endpoints. This is, for instance, the case for the popular logrank test which can be used to compare two survival distributions. Assuming proportional hazards for the survival curves of intervention arm versus control arm with unknown hazard ratio of λ, the stage-one logrank test statistic $X_1 \sim \mathcal{N}\big(-\log(\lambda)\sqrt{\eta_1\, n_1/4}, 1\big)$ is asymptotically normally distributed (Schoenfeld, 1981). Here, η_1 is the fraction of individuals experiencing an event in stage one. Similarly in stage two, $X_2 \,|\, n_2(x_1) > 0, X_1 = x_1 \sim \mathcal{N}(-\log(\lambda)\sqrt{\eta_2\, n_2(x_1)/4}, 1)$. The optimal two-stage test can then be derived by solving the analogous optimisation problem to the one given by (7.1)-(7.3).

All ideas regarding optimisation under uncertainty discussed in Chapter 3 may also be transferred to the continuous setting. An example application for a single-arm trial with continuous endpoint is described in Section 13.

Part III.

Results

8. Examples: Optimal Two-Stage Designs

8.1. Generic optimal two-stage design

Consider the example situation described in Section 1.3.1 or a single-arm design for a response rate under TAU of $p_0 = 0.2$, a point alternative of $p_{alt} = 0.4$, a maximal type one error rate of 5%, and a target power of 80%. Figure 8.1 compares Simon's optimal design with the generic two-stage solution under the additional constraints discussed in Section 2.3.1 but without unimodality constraint (optimal) and the optimal design with unimodality constraint (optimal, unimodal).

The increased flexibility of the generic optimal two-stage design highlights features of the objective function (here expected sample size under $p_0 = 0.2$) that are lost by restricting the solution to group-sequential designs. The sample size function's shape of the generic two-stage solution clearly reflects the fact that the objective targets small expected sample sizes under the null. The stopping-for-futility region is as large as under Simon's design since aggressive early stopping for futility is the most effective way of reducing the expected sample size under the null. On the continuation region, the sample size is mostly increasing in the number of observed responses. This phenomenon was previously criticised by Banerjee *et al.* (2006):

> *'[…A] counter-intuitive feature of this design is that as [x_1] increases, the second-stage sample size, n_2 increases till a certain point and then abruptly becomes zero […].'*

In fact, it is only counter-intuitive from a *conditional* (on $X_1 = x_1$) view: Since larger x_1 imply more stage-one evidence against the null hypothesis, one might expect the required sample size for the second stage to be a *decreasing* function in x_1 on the continuation region as it is the case for the design using an adaptive recalculation heuristic based on conditional power (cf. Section 1.4.1). From an *unconditional* perspective, however, it makes sense that the design reduces the sample size for x_1 values close to the early-futility boundary and compensates by increasing the sample size for interim results that are unlikely to occur under $p = p_0$. The phenomenon is thus a direct consequence of the choice of objective function. A benefit of studying optimal generic two-stage designs is that they reveal features of the objective function that cannot be seen in the corresponding group-sequential solution since it is too heavily regularised (locally constant $n_2(\cdot)$ on the continuation region). These considerations also imply that a global monotonicity constraint as suggested by Shan *et al.* (2016) is not justified, especially not one that is always imposing a decreasing sample size function on the continuation region.

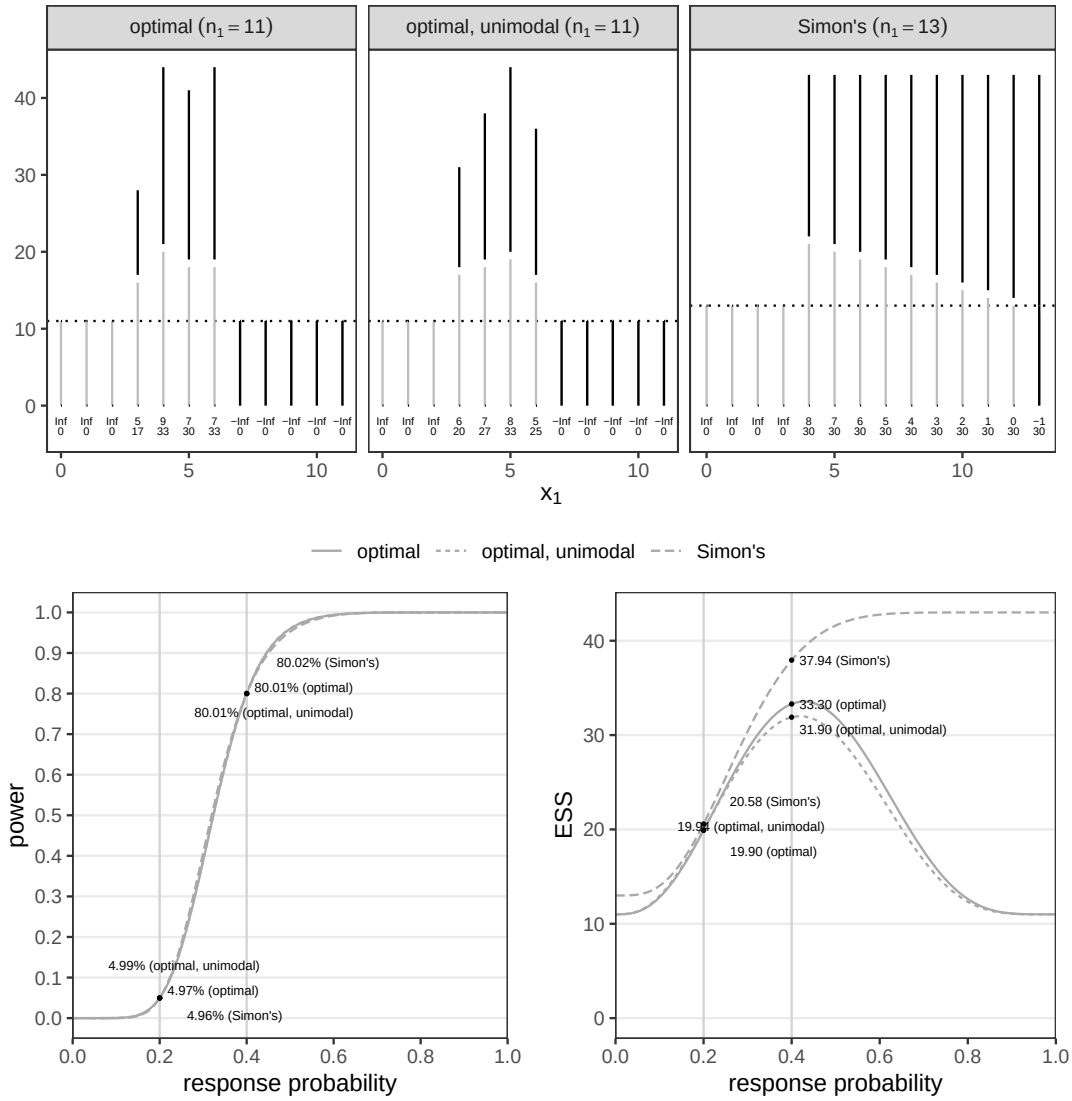

Figure 8.1.: Simon's optimal design compared with the generic optimal two-stage design for the situation described in Section 1.3.1 (optimal), and the corresponding optimal unimodal design (see Section 2.3.1).

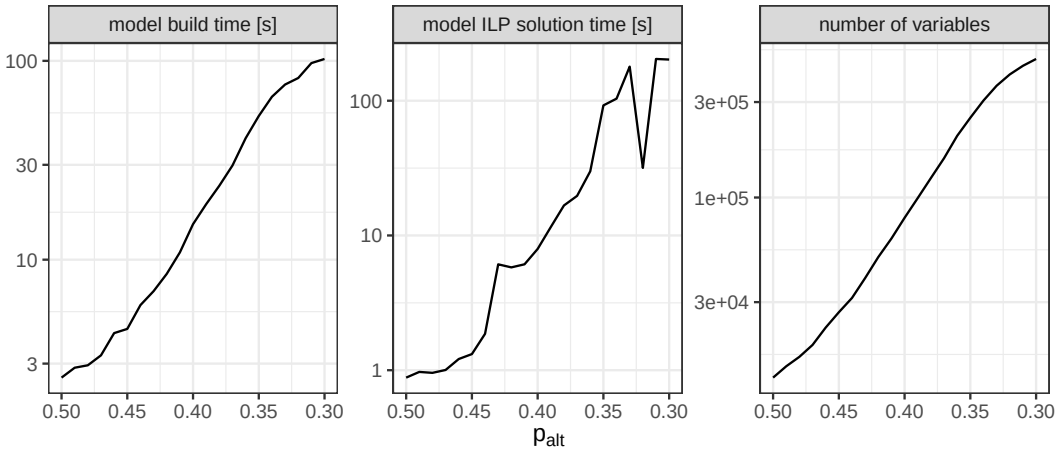

Figure 8.2.: Problem size, problem generation time, and ILP solution time for $p_0 = 0.2$ and varying p_{alt} with $\alpha = 0.05$ and $\beta = 0.2$. Vertical axis is log-10 transformed but axis labels are given on original scale.

In this particular situation, the optimal design without unimodality constraint (optimal, raw) exhibits a non-unimodal sample size function. The restriction to unimodal sample size functions (optimal) has negligible impact on the overall performance in terms of expected sample size under p_0 (increase from 19.90 to 19.94) and is more flexible than a monotonicity constraint. Again interpreting two-stage designs as trial-internally randomised tests, it is not surprising that the more flexible generic two-stage design matches the error rate constraints even better than Simon's design (cf. Figure 8.1). Besides the obvious benefits of allowing early stopping for efficacy, this better exploitation of the permissible error rates explains the reduction in the objective criterion of expected sample size under the null hypothesis (cf. Figure 8.1). The optimal design dominates Simon's design for almost all values of p except for a slight disadvantage on the interior of the null hypothesis due to a larger stage-one sample size. This is caused by the fact that the objective criterion exclusively considers expected sample size on the *boundary* of the null hypothesis and ignores performance on the interior.

To illustrate the efficiency of the proposed solution method, Figure 8.2 shows the total number of binary variables, the time spent on problem generation, and the actual ILP solution time over a variety of problem sizes given by $p_{alt} - p_0$ (smaller differences require larger models). GLPK v4.64 (GNU Project, 2020) was used as ILP solver on a MacBook Pro 2019 with 2.3 GHz Intel Core i9 and 16 GB 2400 MHz DDR4. All regularity constraints discussed in Section 2.3.1 were used except global unimodality of the sample size function. Furthermore, n_{max} was restricted to 1.5 times the required sample size of the one-stage test which is still substantially larger than the maximal sample size considered by Englert *et al.* (2013). Both the problem size (number of variables) and the solution time scale exponentially in the difference between p_0 and p_{alt}. However, even the smallest difference of $p_{alt} - p_0 = 0.1$ is still feasible with $n_{max} = 195$ and an overall solution time of less than 200 seconds. Problem sizes in excess of $n_{max} = 200$ are unlikely to occur in an early phase II trial. In terms of absolute solution time, the

novel approach is fast compared to previous implementations. For $p_{alt} - p_0 \geq 0.2$, the overall solution time (problem generation and solution) is less than 15 seconds. Note that this is the overall time spent including optimisation over a wide range of n_1 values. A comparison with previous methods in term of solution time is difficult since the code is typically not available. Englert *et al.* (2013) provide a script for the statistical computing environment R (R Core Team, 2019) that can be used to compute solutions conditional on individual n_1 values. Due to the exhaustive nature of their custom Branch & Bound algorithm and the implementation as naïve recursion in R, even the solution for individual n_1 values may take several minutes. For instance, the ILP-based method requires 118 seconds to find the overall optimal solution for $p_{alt} = 0.35$. For comparable settings, $n_{max} = 95$ and the $n_1 = 19$ the implementation of Englert *et al.* (2013) could not solve the problem within two hours. Note that this is only conditional on $n_1 = 19$ and the search space considered by the ILP implementation is $n_1 = 19, \ldots, 49$ in this case. Depending on the number of n_1 values for the necessary grid-search overall solution of problems of comparable size could take hours or days. For medium to large problems, the ILP approach is thus orders of magnitude faster and allows to raise n_{max} to values that are much less likely to be binding for most practical problems in early clinical oncology.

8.2. Alternative objective functions

To illustrate the importance of the choice of objective function, the generic optimal two-stage designs minimising either expected sample size under the null hypothesis, under the alternative, or the maximal sample size are computed for the previously introduced example situation (see Section 1.3.1). The minimax design was made unique by combining the minimax objective with expected sample size under p_0 as outlined in Section 2.4. The design is thus the minimax design with smallest expected sample size under p_0. All three designs are compared in Figure 8.3. The minimax design closely resembles a group-sequential one. There is, however, minimal variation in the stage-two sample size caused by the discreteness of the problem. The generic two-stage minimax design has a maximal sample size of 32 which is one less than the optimal group-sequential minimax design reported by Simon (1989) and exactly the same as the single-stage randomised test (see Figure 1.1). The design minimising expected sample size under $p = p_{alt}$ (alternative-design) is markedly different from the one minimising expected sample size on the boundary of $\mathcal{H}_0$. The shape of the sample size function is exactly reverse and *decreasing* on the continuation region. The null-design has the largest maximal sample size of all three designs. This can be explained by the fact that the objective criterion favours aggressive early stopping for futility and small sample sizes for values of x_1 that are likely under $p_0 < p < p_{alt}$. To achieve the overall desired power level, the design must compensate by ensuring a high power and thus large sample size for x_1-values that are less likely under p_{alt}, i.e. towards the stopping-for-efficacy boundary.

This example demonstrates that the choice of objective function is absolutely crucial. All three designs can rightfully claim to be 'optimal' - only that the respective underlying assumptions are completely different. Minimising expected sample size

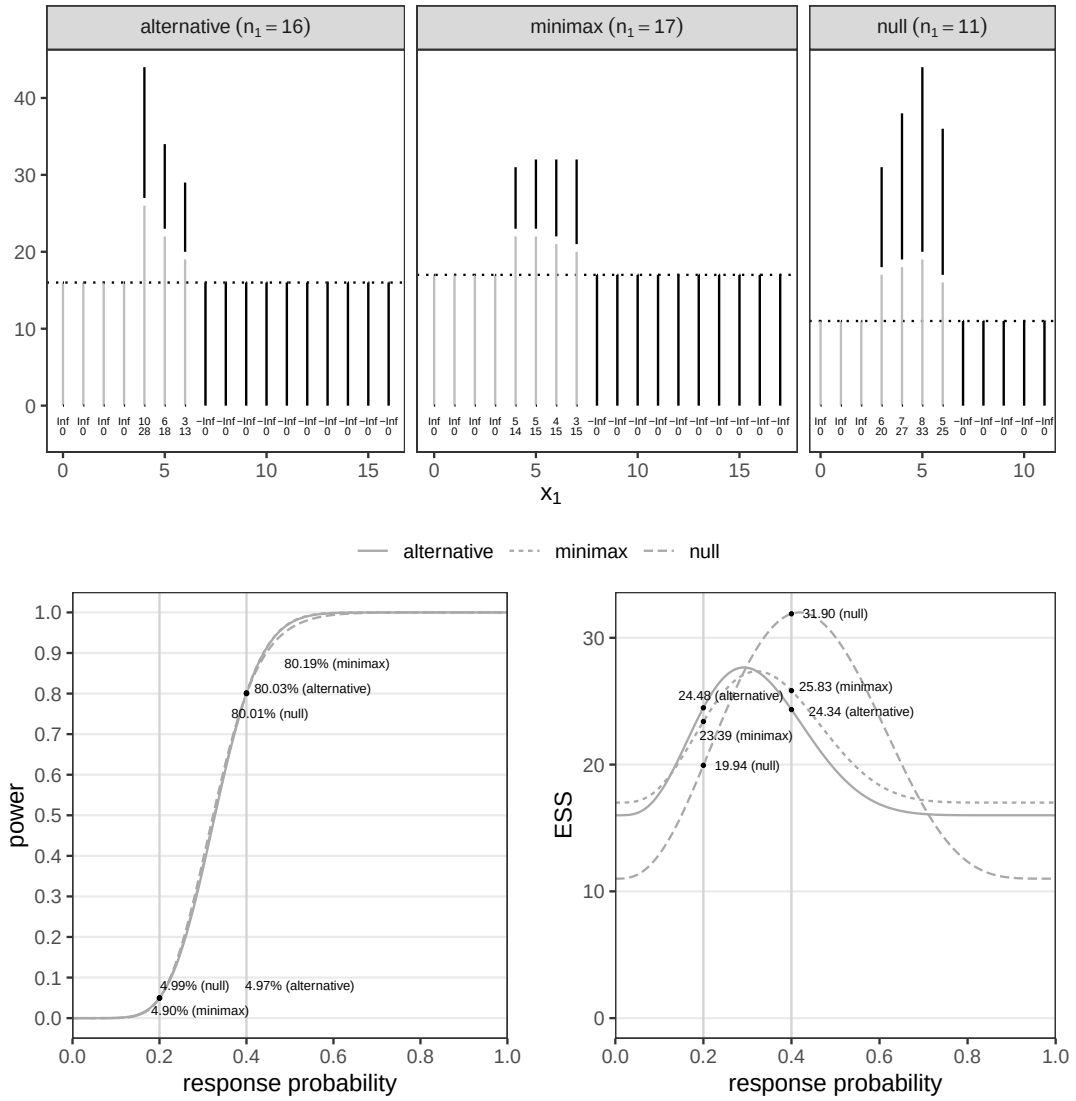

Figure 8.3.: Optimal designs minimising the maximal sample size (minimax), the expected sample size on the boundary of the null hypothesis (null), or the expected sample size under the point alternative (alternative).

under p_0 is sensible when there is severe doubt about the improvement in terms of response probability of a new treatment and trial cost and duration are less important to the sponsor should the new treatment eventually turn out to be effective. The increase in expected sample size under the alternative over the other two designs is, however, substantial. Expected sample size under the alternative is suitable in exactly the opposite situation where the new treatment can be expected to be beneficial. Both of these approaches trade off a reduction in expected sample size at a particular response rate with an increase in the variability and the maximum of their sample sizes. The minimax approach addresses this by directly minimising the worst case sample size but is less effective than the respective optimal designs for the particular response probabilities that they are optimised for. Note that the minimax design is not a one-stage design due to the discreetness of the test statistic. By exploiting early stopping for efficacy and futility, the permissible error rates can be exhausted more closely than with a one-stage test. In practice, it is hard to decide which of the respective assumptions is most valid and an approach for continuously interpolating between the different objectives on a principled basis could help investigators to pick designs that more honestly reflect their uncertain prior knowledge about the anticipated response probability. Chapter 3 discusses ways of addressing this problem by quantifying *a priori* uncertainty about p directly.

9. Examples: Optimisation Under Uncertainty

9.1. Prior choice

Again consider the example first introduced in Section 1.3.1. The response rate under TAU is $p_0 = 0.2$. Assume that the initial assumptions about the true response rate under the new treatment can be summarised in a Beta prior with mean 0.35 and standard deviation 0.1. Using equations (3.8) and (3.9) yields a $\mathrm{Beta}(7.61, 14.14)$ distribution. Further assume that data from a previous dose finding study is available and 4 out of 10 subjects showed a response under similar overall conditions and a comparable definition of tumour response. The maximum likelihood estimate of the response rate is 0.4 and consistent with the previously assumed $p_{\text{alt}} = 0.4$.

Following the proposed procedure for updating a uniform prior with the phase I data, the phase I posterior is a $\mathrm{Beta}(11.61, 20.14)$ distribution. Due to slight differences in the treatment procedures between phase I and the planned phase II trial, a robustification with $\epsilon = 0.2$ is deemed appropriate. Finally, $\overline{p} = 0.7$ is imposed to prevent excessive weight on very large response rates. Without conditioning on $p \leq 0.7$ the *a priori* probability of a response rate larger than 0.7 would still be more than 6%. The steps of constructing the final 'pragmatic' prior $\varphi = \varphi_{7.61,14.14,0.2}|_{\leq 0.7}$ are visualised in Figure 9.1 (see Section 3.2). The combined effect of robustification and conditioning on $p \leq 0.7$ is clearly visible by the fact that the lower tail of the prior is lifted to a minimum of $\epsilon = 0.2$ whereas the upper tail is 0. The impact of a robustification is more pronounced in situations with a more concentrated prior since the tails of the Beta distribution are quickly approaching zero even for moderately large a, b. Overall, the approach to defining a pragmatic prior proposed here is flexible enough to represent a wide range of situations without being overly complicated. Of course, it can be extended to encompass multiple informative mixture components from different phase I studies but a practical benefit of adopting a multi-modal prior is questionable. The key elements are an honest representation in terms of location and scale of available *a priori* information and the option to robustify and restrict the prior to a plausible range.

The resulting optimal designs under the non-informative and the final 'pragmatic' prior are compared in Figure 9.2. The non-informative design closely resembles the minimax design (see Section 8.2, Figure 8.3). The maximal sample size is, however, slightly larger (36 vs. 32 for the minimax design) and the stage-one sample size is smaller (9 vs. 17). This is due to the fact that under a uniform prior, the expected sample size reduces to the average sample size (all x_1 are equally likely during the interim analysis) putting more weight on the stage-one sample size while the minimax

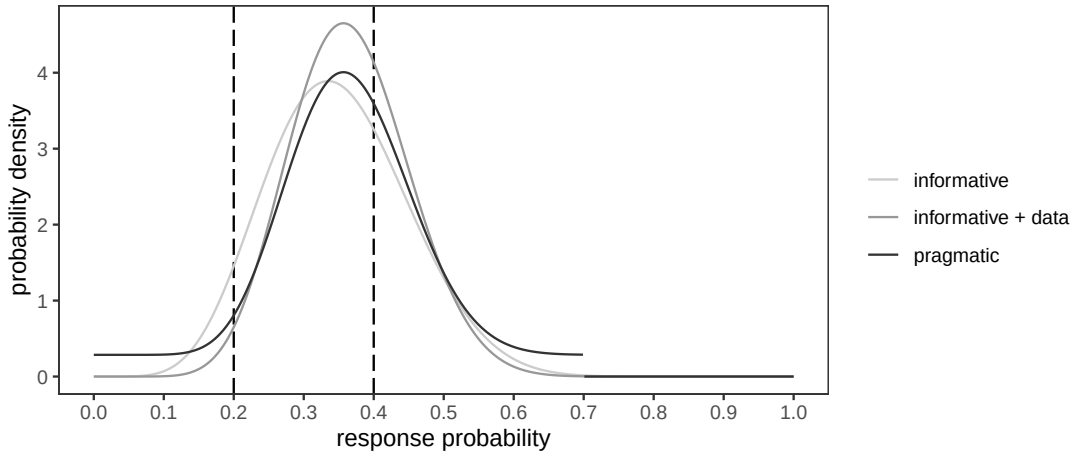

Figure 9.1.: Individual step of prior construction: 1) start with a maximum-entropy non-informative prior (here: uniform distribution) 2) update with phase I data (informative) 3) make prior robust by 'mixing in' a non-informative prior again (robust) 4) restrict prior on maximal plausible range (pragmatic).

criterion ignores the stage-one sample size entirely. Consequently, group-sequential designs can be interpreted as being (near) optimal with respect to both maximal as well as average sample size (under a non-informative prior). Generic response adaptive designs may thus only yield advantages over group-sequential ones under informative planning priors.

Under the pragmatic planning prior the optimal design becomes more flexible but the sample size remains less variable than that of designs minimising expected sample size conditional on a single response rate (cf. Figure 8.3). From a non-Bayesian perspective, this can be interpreted as regularising the objective criterion of expected sample size by averaging over a range of values of p where the weight function is given by the respective prior density. Yet, the Bayesian view offers richer insights into the validity of different objective criteria and gives a formal framework for eliciting the weight function (i.e. the prior). Overall, a Bayesian approach is more principled than simply minimising expected sample size under a single response probability, easier to adapt to a specific situation, and more practical since a reduced variability of the stage-two sample size and a smaller maximal sample size (i.e. less 'overfitting' of a particular response rate) makes the resulting designs easier to conduct.

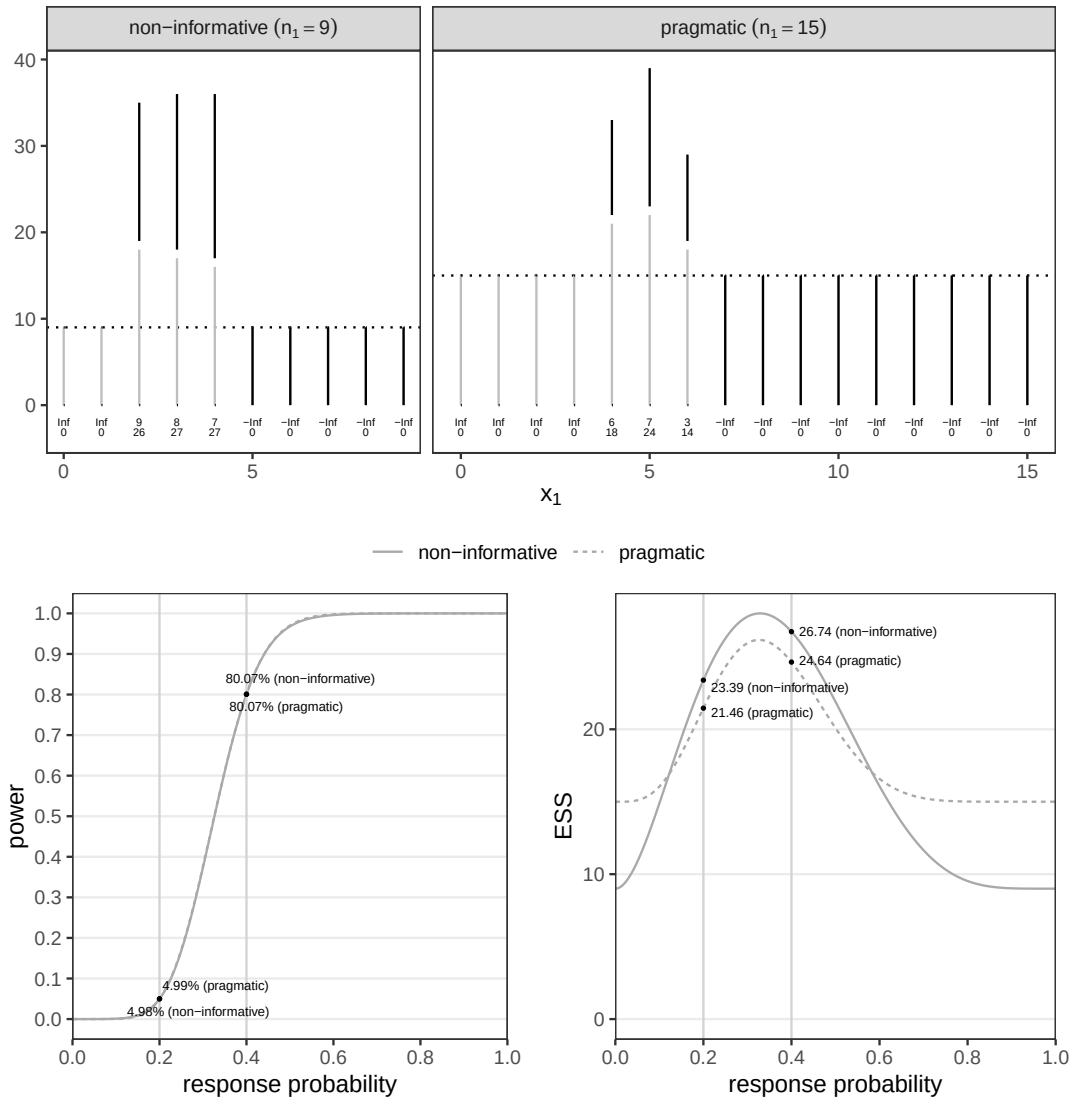

Figure 9.2.: Comparison of optimal designs minimising expected sample size under non-informative and pragmatic priors as show in Figure 9.1.

9.2. Bayesian power constraints

Assume that the 'pragmatic' prior $\varphi = \varphi_{11.61,20.14,0.2}|_{\leq 0.7}$ derived in Section 9.1 is considered appropriate (see Figure 9.1). Let the minimal clinically relevant effect be $\overline{p} = 0.3 = p_0 + 0.1$. Furthermore, assume that an *a priori* probability of $\gamma = 2/3$ for exceeding the power threshold of $1-\beta$ is deemed acceptable. The designs minimising expected sample size under this prior under power constraints on $p_{\text{alt}} = 0.4$, $p_{\text{alt}} = 0.36$ (1/3 quantile of the conditional prior), and under an expected power constraint are shown in Figure 9.3.

Evidently, the quantile-approach requires a larger design than the one under $p = 0.4$ since the conditional quantile is smaller ($\approx 0.36 < 0.4$). This relation, however, critically depends on the desired *a priori* assurance for exceeding the power threshold. For very large γ, the conditional $(1-\gamma)$-quantile of the prior converges to p_{MCR}. The quantile approach thus merely shifts the problem from having to specify p_{alt} directly to selecting an *a priori* assurance level γ. Furthermore, its practical applicability is limited by the fact that it requires a solid understanding of two conditional probabilities ($1-\beta$ and γ) with related but different roles in defining the power constraint. As a theoretical exercise, however, it illustrates a principled way of eliciting p_{alt} that is firmly rooted in Bayesian theory and incorporates a clear distinction between the minimal clinically relevant response probability p_{MCR} and the *a priori* relative likelihood of different response probabilities p.

An expected power constraint is an attractive alternative since it does not require the specification of an additional assurance parameter γ. Expected power is a functional of the entire power curve for $p \geq p_{\text{MCR}}$ and can be sensitive to the tail behaviour of the chosen prior. To understand why this is the case, consider the trade-off between power at $p_2 = 0.42$, the prior mean conditional on $p \geq p_{\text{MCR}}$, and $p_1 = 0.7$, the upper boundary of the plausible range for response probabilities. The trade-off between the two under expected power is governed by

$$\mathrm{d}\,\mathrm{power}(p_2) \approx -8.56 \cdot \mathrm{d}\,\mathrm{power}(p_1). \tag{9.1}$$

A decrease in power of one percent point at p_1 (*a priori* likely response probability) can be thus be compensated for by increasing power at p_2 by 8.56 percent points. During optimisation, this trade-off reduces the pressure to increase power at the centre of mass of the prior since a very high power (close to 100%) for $p > 0.6$ can already be achieved with very low sample sizes (rate difference greater than 0.4 from p_0). A prior with heavy upper tails will thus tend to result in smaller-than-expected designs and great care should be taken to ensure that the upper tails of φ adequately reflect the available *a priori* information. If p_2 was larger than $\overline{p}$, $\varphi(p_2) = 0$ and a reduction in power at p_1 could no longer be compensated by an increase in power at p_1 since $\varphi(p_1)/\varphi(p_2) = \infty$. The impact of this 'unrealistic trade-off' phenomenon is increasing in the difference $\overline{p} - p_0$ since this governs the non-zero upper tail area of the prior. To illustrate how the parameter $\overline{p}$ affects the resulting design, Figure 9.4 compares optimal designs under $\overline{p} = 0.5, 0.7$ and 1.0.

As the prior cutoff approaches p_{MCR}, the prior mass shifts towards lower response probabilities and the size of the optimal design increases since high power at very

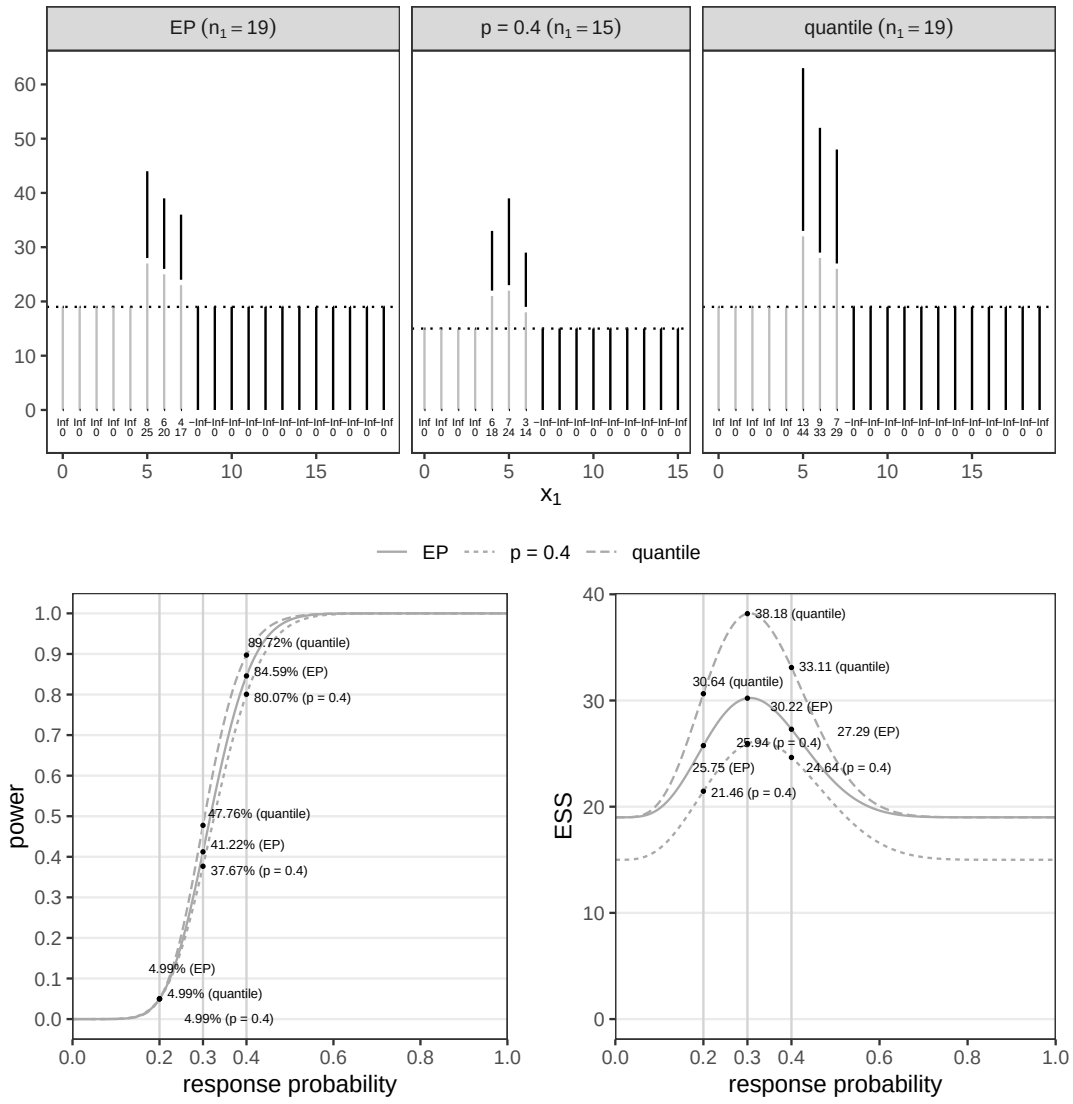

Figure 9.3.: Expected power constraint (EP), power constraint on the observed phase I rate (4/10), and on the 33% quantile of the prior distribution conditional on a relevant effect (quantile).

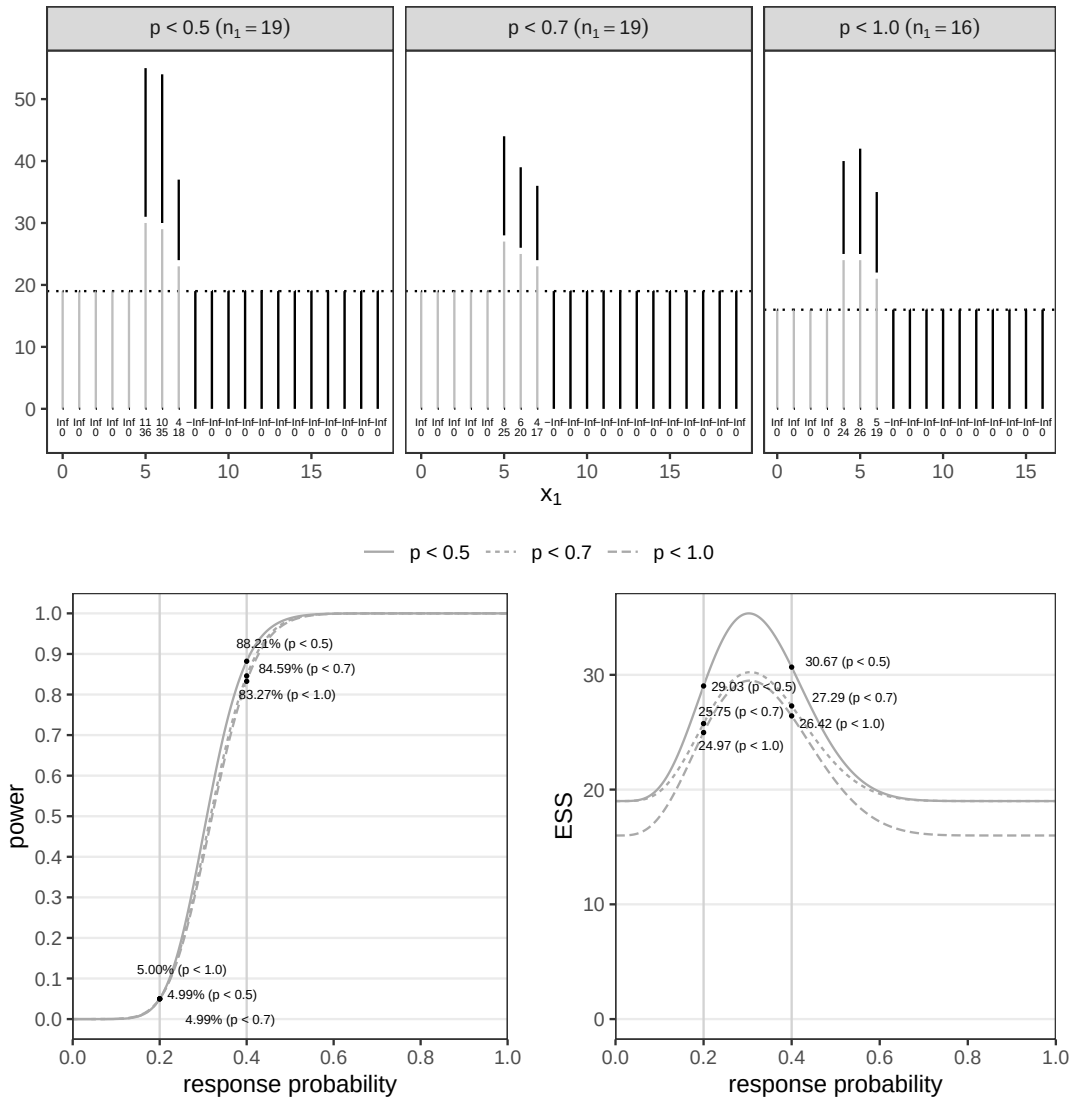

Figure 9.4.: Effect of the upper-tail cutoff for the pragmatic prior on the resulting optimal design under an expected power constraint.

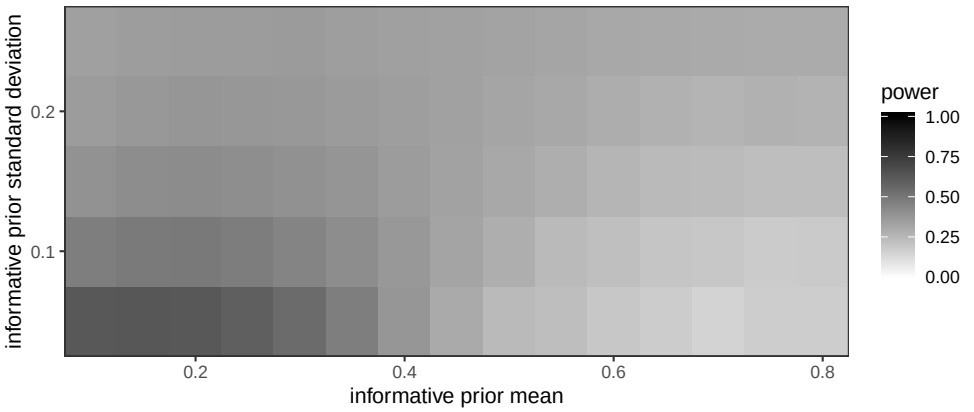

Figure 9.5.: Surface plot of the power at $p_{\text{MCR}} = 0.3$ for different means and standard deviations of the initial informative prior component of the pragmatic prior.

large response rates can no longer compensate for a reduction at lower values of p. *Vice versa*, liberal upper tails together with prior robustification can lead to smaller-than-expected designs since the power is almost one in the upper tails of the prior.

This example demonstrates how difficult it is to give a single definition of a power constraint under uncertainty about p that properly reflects real world objectives and *a priori* information. Seemingly simple approaches can be hard to justify rigorously and might lead to counter-intuitive results. In practice, Bayesian power constraints are only relevant in situations where the prior reflects an *a priori* conviction that the true effect might be substantially larger than the minimal clinically relevant one. It would then be inefficient to simply power on p_{MCR}. Expected power is particularly appealing since it requires only the specification of p_{MCR} (which might equal p_0) and the prior density φ. The example, however, also shows that even a design with a sufficient expected power might have a surprisingly low chance of being sufficiently powered. Whichever approach is preferred, a clear distinction between the minimal clinically relevant response probability and arguments based on the relative *a priori* likelihood of response probabilities is only possible within the Bayesian framework. Transparently communicating the power on the entire range of relevant response probabilities is then key to deciding whether a particular power constraint is compatible with the trial objectives.

Figure 9.5 shows the effect of different choices of the initial mean and standard deviation of the informative prior component on the power of the corresponding optimal design at the minimal clinically relevant response probability. For small prior standard deviations (high *a priori* certainty) the power at p_{MCR} depends critically on the prior mean since the expected power constraint is then most similar to simply calculating the power on the prior mean. If relevant effects are likely to be small (small prior standard deviation and small prior mean), the required sample sizes are large. *Vice versa* large certainty in large relevant effects leads to low power at p_{MCR} and consequently smaller trials. As the prior standard deviation increases the dependency of the power at p_{MCR} on the prior mean is reduced since the prior approaches a uniform distribution on $[\,0, 0.7\,]$.

9.3. Utility maximisation

A major drawback of utility-based methods as described in Section 3.4 is the need to define the utility function. The problem reduces to selecting the values of $\lambda_{+|-}$ and $\lambda_{+|+}$ for the utility score defined in equation 3.27. Data on expected future payout tends to be highly situational and uncertain. However, the utility optimisation approach can also be employed to reveal implicit utility assumptions of any given design. For instance, one could strive to find the pair $\lambda_{+|-}$ and $\lambda_{+|+}$ that would make a particular design $\mathcal{D}$ rational (i.e., optimal) under the corresponding s_u. To this end, assume that a design $\mathcal{D}_0$ is given that has been optimised for expected sample size under a prior φ and potentially unknown constraints on type one and type two error rates. One can then try to find the combination of $\lambda_{+|-}$ and $\lambda_{+|+}$ under which the operating characteristics of $\mathcal{D}_0$ (i.e. its power curve) are close to the corresponding optimal design under $\mathrm{s}_u(\lambda_{+|-}, \lambda_{+|+})$. This approach can serve as a plausibility check of any given design since it is typically easier to decide whether a particular combination of $\lambda_{+|-}$ and $\lambda_{+|+}$ is plausible than to fix exact values upfront. To formalise this idea, assume that the 'closeness' of power curves can be measured by their maximal absolute difference

$$\delta\left(\ \mathrm{power}_{\mathcal{D}_1}, \mathrm{power}_{\mathcal{D}_2}\ \right) := \sup_p \left|\ \mathrm{power}_{\mathcal{D}_1}(p) - \mathrm{power}_{\mathcal{D}_2}(p)\ \right|. \tag{9.2}$$

Furthermore, let $\mathcal{D}^*_{\mathrm{s}_u}(\lambda_{+|-}, \lambda_{+|+})$ be the minimiser of s_u for given $\lambda_{+|-}$, and $\lambda_{+|+}$. The implicit utility parameters $\lambda_{+|-}(\mathcal{D}_0)$ and $\lambda_{+|+}(\mathcal{D}_0)$ are then given as the solution of

$$\underset{\lambda_{+|-},\,\lambda_{+|+}}{\mathrm{argmin}} : \delta\left(\ \mathrm{power}_{\mathcal{D}_0},\ \mathrm{power}_{\mathcal{D}^*_{\mathrm{s}_u}(\lambda_{+|-},\lambda_{+|+})}\ \right). \tag{9.3}$$

Due to the discreteness of the problem of finding $\mathcal{D}^*_{\mathrm{s}_u}(\lambda_{+|-},\ \lambda_{+|+})$ this objective function is not differentiable. In practice, an approximate local minimum can be obtained by heuristically exploring some combinations of $\lambda_{+|-}$ and $\lambda_{+|+}$ to get an initial guess with roughly the same power curve as under $\mathcal{D}_0$ before fine-tuning the parameters using a derivative free local optimiser like the Nelder-Mead algorithm (Nelder *et al.*, 1965). This typically requires at least tens if not hundreds of function evaluations of the objective function and thus the global optimisation of an equal number of utility-maximisation problems. Without the efficient solution strategy developed in Section 2.3 this would be utterly hopeless.

Consider, for $\mathcal{D}_0$, the design minimising expected sample size subject to the expected power constraint discussed in Section 3.3 and Figure 3.3. The first step to finding a matching utility-maximising design for the score given in equation (3.27) is an initial grid search over a range of plausible values for $\lambda_{+|-}$ and $\lambda_{+|+}$. A rectangular grid is unlikely to be effective since it is the relative size of the two parameters that governs the ratio of error rates and the average magnitude of both the steepness of the power curve. An initial grid together with the attained values of δ and the minimiser on this grid is shown in the left panel of Figure 9.6. Since δ is increasing towards the boundaries of this search pattern it is reasonable to assume that the grid sufficiently explores the overall space. The crude initial guess can then be refined by using a local derivative-free optimiser. Here, the Nelder-Mead algorithm implemented in the

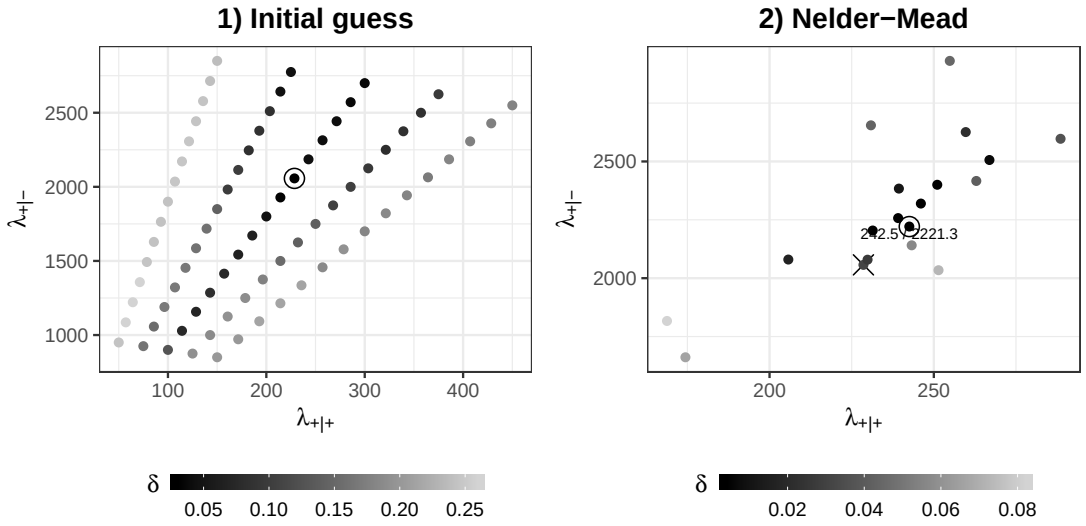

Figure 9.6.: Initial grid-search evaluation of the δ function 1) and trace plot of the parameter combinations explored during the local Nelder-Mead refinement 2); In both plots, the circle indicates the best fitting parameter configuration and in 2) the cross indicates the starting point of the local optimisation (result of grid search in 1).

base R package is employed (Nelder *et al.*, 1965; R Core Team, 2019). The initial guess (cross), the trace of points visited during the optimisation procedure, and the corresponding optimum found upon convergence with an absolute tolerance of less than 0.0001 (circle) are shown on the right panel of Figure 9.6. The matched utility parameters are $\lambda_{+|+} \approx 243$ and $\lambda_{+|-} \approx 2221$. The utility-maximising design $\mathcal{D}^*_{\mathrm{s}_u}(243, 2221)$ for the matched utility parameters is compared with the initial design $\mathcal{D}_0$ in Figure 9.7.

Per-patient costs in oncological clinical trials are high (Battelle Technology Partnership Practice, 2015). A conservative guess is a value in the order of magnitude of 100 000 US$ (overall: treatment, fees, follow-up). This directly translates to an implicit risk weighted future benefit upon *successful* rejection of the phase II null hypothesis of 24.3 million US$ and an implicit cost of 222.1 million US$ upon *wrongful* rejection of $\mathcal{H}_0$ (costs of conducting futile phase III). Unconditional success rates of substances (likelihood of approval, LOA) from phase I to approval range from about 9% to 12% (Thomas *et al.*, 2016). Assuming that substances with promising phase I results (4/10) and a successful phase II trial tend to be more successful in in phase III, a success rate of 1 in 3 can be assumed. Since, 24.3 million US$ is the assumed future *risk weighted* profit, the overall anticipated profit from a successful approval under these assumptions is 72.9 million US$.

The overall matched s_u score of the original $\mathcal{D}_0$ is 84.3 (i.e., 8.43 million US$). This figure does not yet include the fixed costs of running the phase II trial. Under the assumed per-patient costs, the expected variable cost of $\mathcal{D}_0$ (expected sample size multiplied with average per-patient costs) is 2.67 million US$. Considering the fact that fixed costs make up a major part of the overall budget for small trials, these figures are plausible given reported phase II trial costs of 5 to 12 million US$ (Sertkaya *et al.*, 2014). If expected utility is directly maximised for the matched paramet-

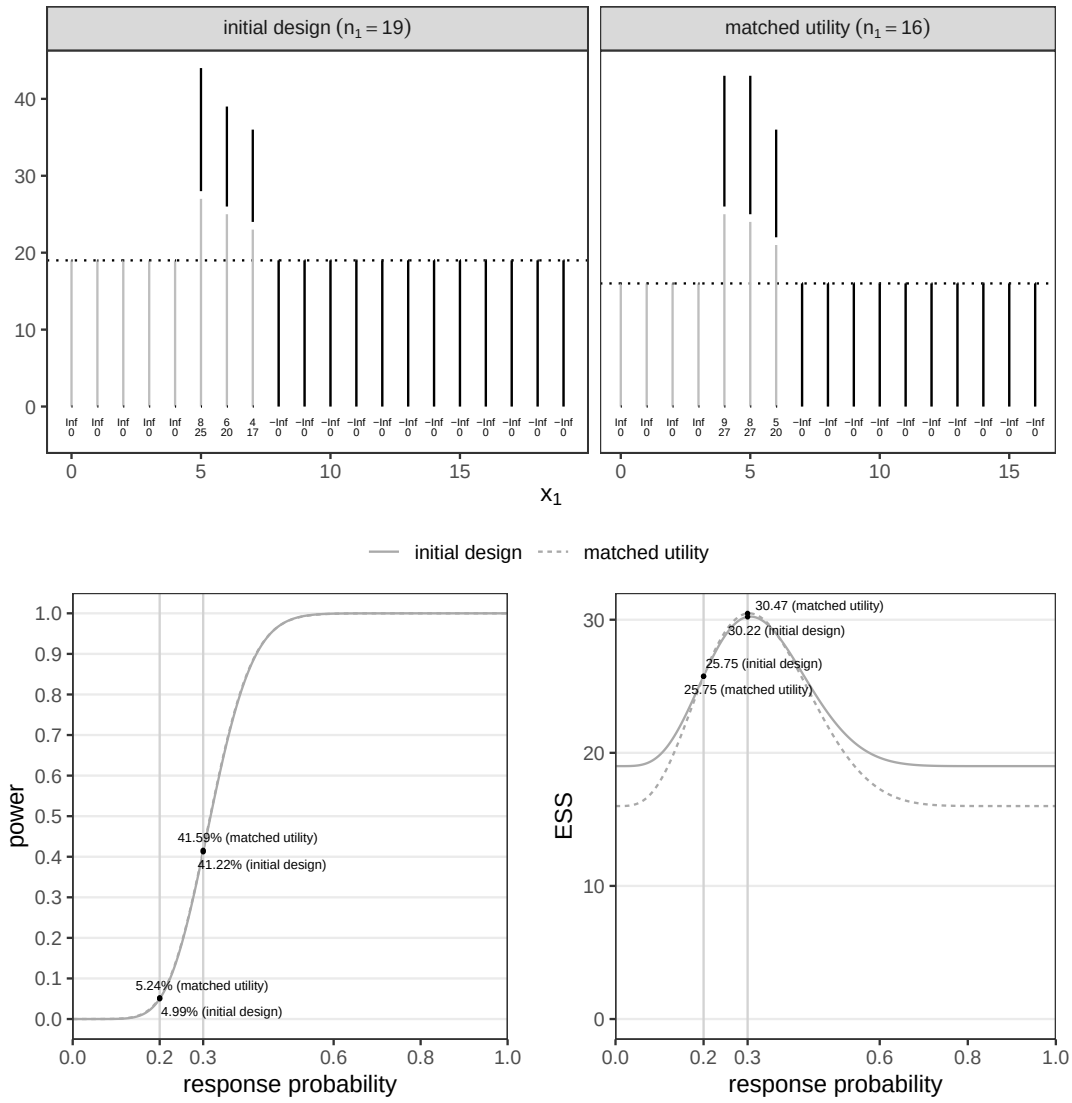

Figure 9.7.: Comparison of the initial design with its matched utility design (see Section 9.3).

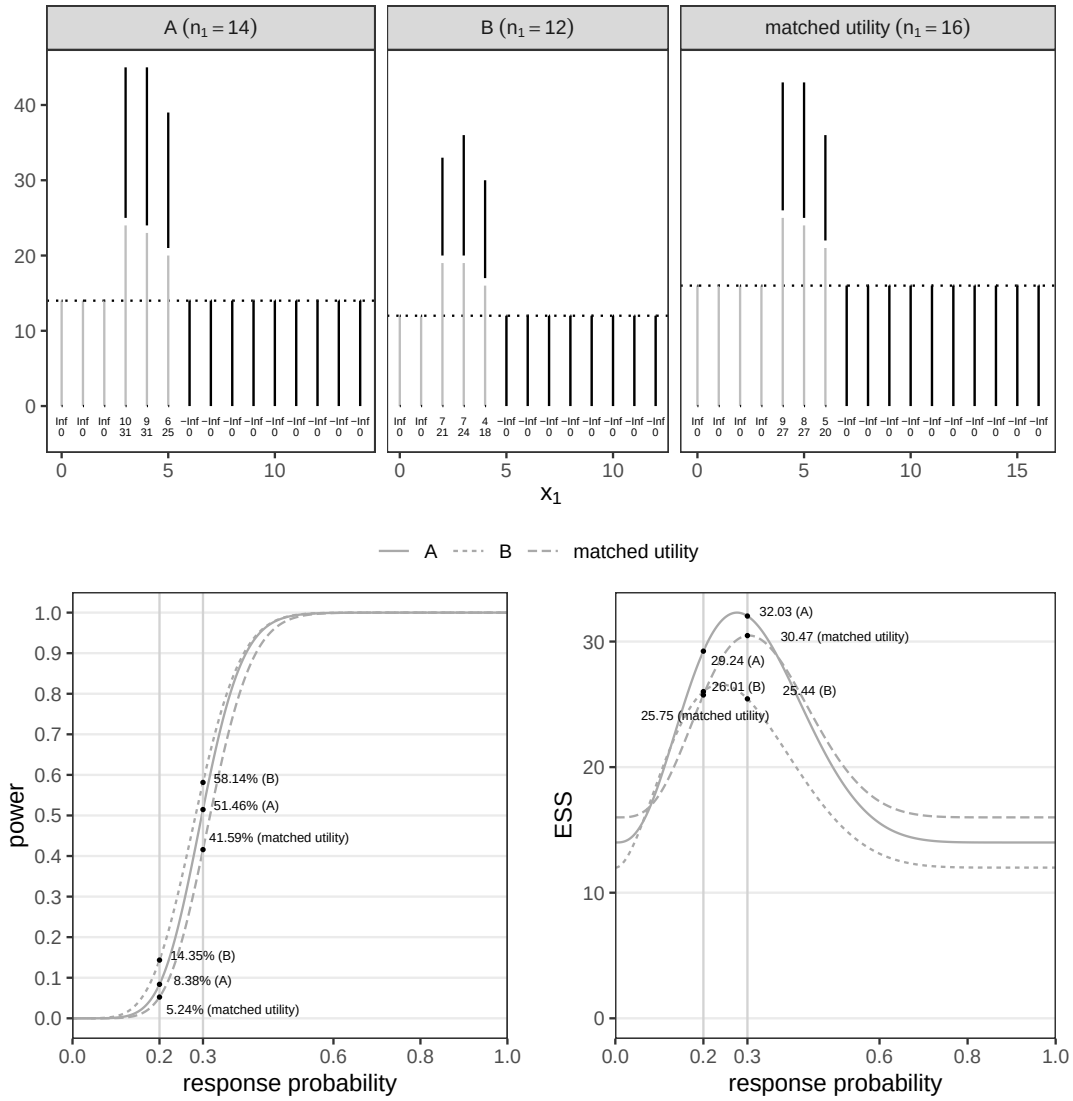

Figure 9.8.: Designs optimising the same utility function under lower reward for correct rejection (A) than the matched utility parameters and under lower penalty for wrong rejection (B) compared with the matched utility design.

ers, it can be increased from 8.43 to 8.72 (under the matched parameters $\lambda_{+|+} \approx 243$ and $\lambda_{+|-} \approx 2221$).

To explore the sensitivity of the utility-maximising design under changes to the assumptions, two variations are considered. For case A, the anticipated (non risk weighted) profit is larger (100 million US$ instead of 72.9 million US$). The corresponding utility parameters are $\lambda_{+|-} \approx 2221$ and $\lambda_{+|+} \approx 333$ (risk weighted profit of 33 million US$). For case B, the costs of a futile failed phase III program are instead reduced from 222.1 million US$ to 150 million US$ ($\lambda_{+|-} \approx 1500$ and $\lambda_{+|+} \approx 243$). Figure 9.8 explores the sensitivity of the utility-optimising design towards these changes of the underlying parameters. In case A, the power of the optimal design increases because the positive incentive upon successful rejection is raised. Note that since there is no strict error rate constraint, this also leads to an increase in type one error

since the costs of a failed phase III program remain fixed and the trade-off between the type one and type two error rates is governed by the relative magnitude of both parameters. In case B the penalty for a wrongful rejection of the null hypothesis is substantially reduced instead. This also leads to an increased type one error rate but also a much higher power even at the minimal clinically relevant response probability. This demonstrates how a more or less arbitrary choice of error rate constraints can be replaced by situation-specific expected-utility-based rational arguments. In practice, though, a precise determination of the required utility parameters is unrealistic and designs with a maximal type-one error rate of more than 10% are rarely realised due to ethical concerns. Still, the utility-based approach can be used to guide the choice of α and β within generally accepted ranges. This should especially be of interest when the optimal error rates are *lower* than the typical $\alpha = 0.05$ and $\beta = 0.2$.

10. Examples: Bayesian Inference

10.1. Posterior distribution and posterior mean estimator

To illustrate the impact of prior choice on Bayesian inference as discussed in Chapter 4, consider the design minimising expected sample size under the 'pragmatic' prior derived in Section 3.1 subject to a minimal expected power of 80% (cf. Section 3.3). Rather than discussing the shape of the posterior distribution for individual outcomes, a comparison in terms of the corresponding posterior mean estimators is presented. Their respective bias, mean absolute error (MAE), and root mean squared error (RMSE) are plotted against the response probability p under the given design in Figure 10.1. Evidently, the performance difference between the three uninformative prior choices are small with the correct Jeffreys prior achieving the smallest point-wise absolute bias. The pragmatic prior, however, is strongly biased towards its centre of mass. This bias allows it to achieve a much better performance in terms of precision in the area of interest roughly between p_0 and $p \approx 0.6$.

The shape of the correct Jeffreys prior itself is interesting. Although the deviation from the naïve Jeffreys prior for a fixed-size binomial experiment is rather small for the situation considered here, a slight 'bulge' in the area of interest is discernible in Figure 10.1. The characteristic behaviour of the Jeffreys prior becomes much clearer when comparing designs minimising expected sample size under different assumptions on the response probability. In Figure 10.2, the previously discussed design and its Jeffreys prior is compared with the design minimising expected sample size under the point null of $p_0 = 0.2$ and a minimal power of 80% at $p_{\text{alt}} = 0.4$. Restricting the plotting area to $[0.05, 0.95]$ reduces the dominance of the tail behaviour of both Jeffreys priors towards 0 and 1. The characteristic deviations from $\text{Beta}(0.5, 0.5)$ caused by the biased sampling scheme are then more clearly discernible. The Jeffreys prior under the pragmatic design prior has a less pronounced 'bulge' due to the smaller overall size of the design but it is located in the actual area of interest between p_0 and $p \approx 0.6$. The null design, however, exhibits a more pronounced shift of mass from $\text{Beta}(0.5, 0.5)$ due to the larger overall size of the design. Furthermore, the prior mass is shifted towards larger response probabilities which might seem counter-intuitive at first. Since the Jeffreys prior is proportional to the square root of the Fisher information, it is increased in areas with large expected sample size (i.e., large (Fisher) information). The pragmatic prior design minimises expected sample size under the unconditional prior, typically leading to an (at least approximately) monotonically decreasing sample size on the continuation region and larger expected sample sizes towards the boundary of the null hypothesis. Consequently, the Jeffreys prior increases the weight in this region as compared to the $\text{Beta}(0.5, 0.5)$ distribution. The situation is entirely reversed when minimising expected sample size under the null.

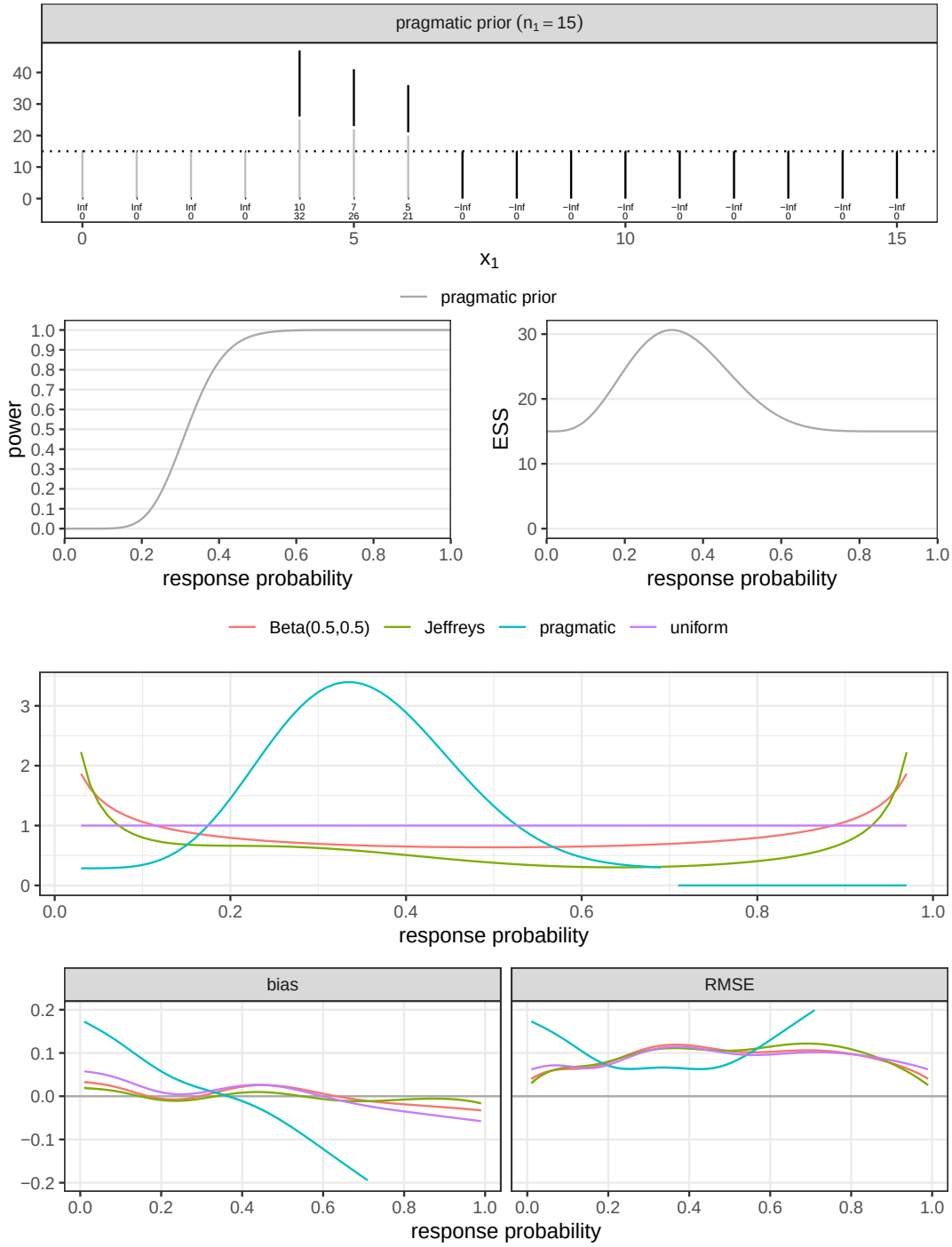

Figure 10.1.: Bias and root mean squared error (RMSE) for the posterior mean estimator with the informative 'pragmatic' prior derived in Section 3.3. The design minimises expected sample size under the same prior and a minimal expected power of 80%. Three uninformative priors (the naïve Jeffreys prior for a fixed-size binomial experiment, i.e., $\mathrm{Beta}(0.5, 0.5)$, the uniform, and the actual design-specific Jeffreys prior) and the informative 'pragmatic' prior are considered.

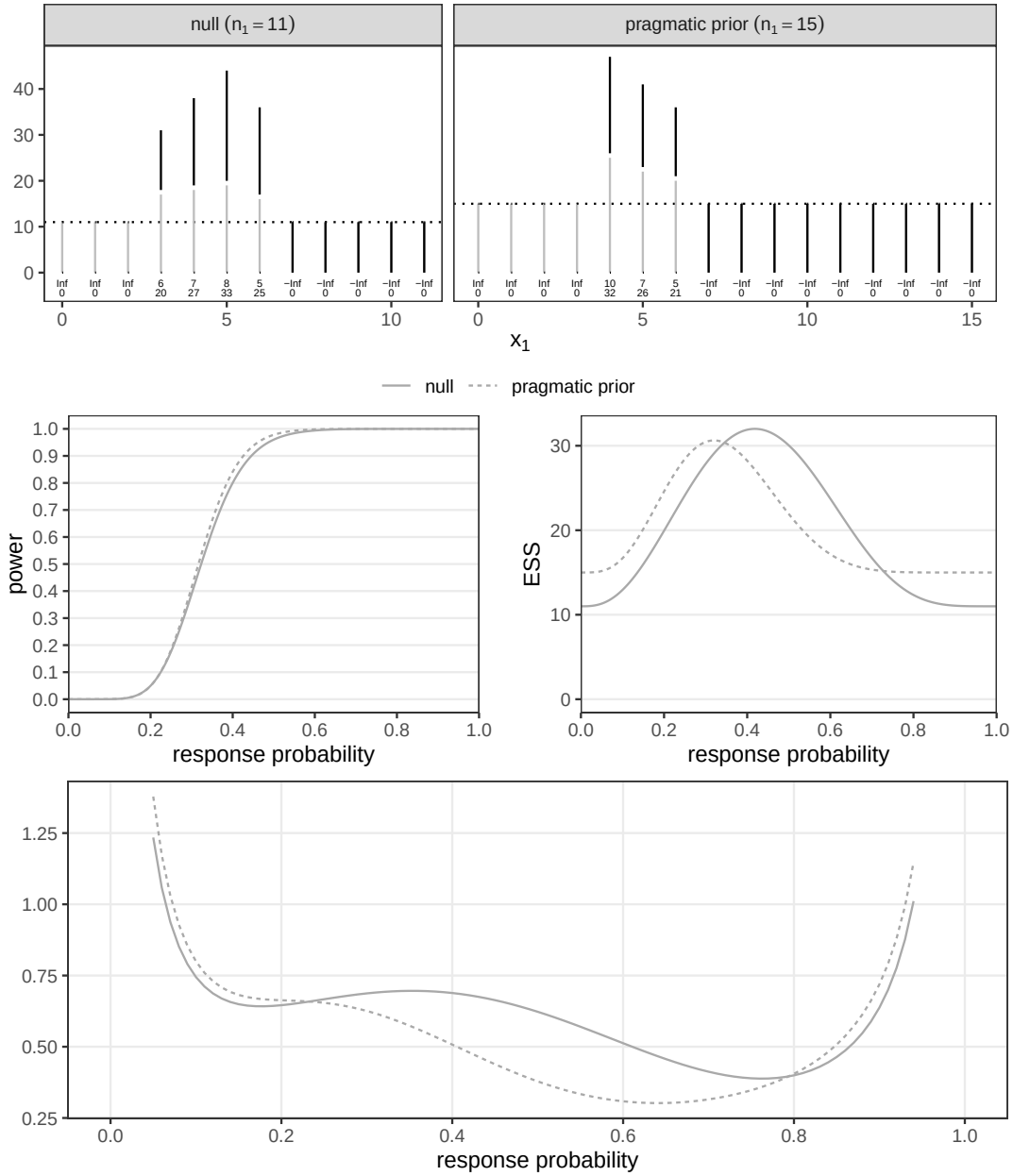

Figure 10.2.: Corresponding Jeffreys priors and design characteristics for the design minimising expected sample size under the pragmatic prior ('pragmatic') and the one minimising expected sample size under the point null of $p_0 = 0.2$, plotting of the prior densities is restricted to $[0.05, 0.95]$ to make the crucial differences more discernible (otherwise dominated by the prior behaviour towards 0 and 1).

This objective criterion induces a shift of expected sample size towards large response probabilities. It follows that the Fisher information is increased for large response probabilities and that the mass of the Jeffreys priors is shifted towards larger values of p. The Jeffreys prior thus tends to imply larger *a priori* probability mass for response probabilities that are unlikely under the planning prior when expected sample size in minimised.

These observations demonstrate the dilemma of objective Bayesian inference. Although any uninformative prior choice reduces the bias of the posterior distribution as measured by the absolute bias of posterior mean estimator, the implied priors are utterly inconsistent with any realistic phase II situation (see also Section 3.1) due to the extremely heavy tails. As outlined above, the Jeffreys prior also tends to increase the relative likelihood of response probabilities where the expected sample size is small. This leads to the slightly paradoxical situation that minimising expected sample size under the null implies higher belief in large alternatives and *vice versa*. Non-informative priors can thus be seen as a valuable tool to reduce bias in the posterior distribution but applied researchers need to be aware of the potentially implied inconsistencies with their original planning assumptions. Early phase II trials in oncology are a situation where, even in the complete absence of compound-specific *a priori* information, experience commands a healthy scepticism about the assumed response rates of a new compound (see Section 3.1).

11. Examples: Frequentist Inference

11.1. Point estimators, p values, and confidence intervals

Consider the 'pragmatic' design minimising expected sample size under the informative prior derived in Section 9.1 subject to a conditional power constraint. First, precision (RMSE) and bias of the frequentist maximum likelihood estimator (MLE) and the unbiased Rao-Blackwellised estimator (RBE) are compared with the posterior mean estimates under the respective Jeffreys priors (PMEJ) and the pragmatic prior (PMEP). The design and the results are shown in Figure 11.1. Firstly, the RBE is indeed unbiased. In this particular situation, the RBE reduces to the stage one maximum likelihood estimator since the design has an injective sample size function on its continuation region. The complete ignorance towards stage-two data means that the RBE is dominated in terms of precision by all other estimators on the region of most interest. The MLE and the PMEJ show very similar performance with slight advantages for the PMEJ, both in terms of (absolute) bias and RMSE on the region of interest. The PMEJ's superior performance can be explained by noting that the MLEs popularity is mostly based on its favourable *asymptotic* properties whereas the PMEJ directly minimises the quadratic Bayes risk for the exact design used, i.e. it prioritises finite-sample properties. Ultimately, however, the PMEJ converges towards the MLE for a single-stage design as the sample size increases. This can easily be seen from the fact that the $\widehat{p}_{\text{PMEJ}}(x_1, x_2) = (x_1 + x_2 + 0.5)/(n_1 + n_2(x_1) + 0.5)$ and

$$\forall\ (x_1, x_2) \text{ with } x_1 + x_2 = x \leq n: \quad \frac{x + 0.5}{n + 0.5} \to \frac{x}{n} \quad \text{as } n \to \infty \tag{11.1}$$

indicating that the two estimators are not too different. The differences between 'objective' Bayesian inference and the use of the informative planning prior were discussed in Section 10.1 and the PMEP is only included in the comparison for reference.

Figure 11.2 depicts the density of the resulting p values (top panel) and the corresponding cumulative distribution functions of the p values under the null and a point alternative ($p = 0.4$). In line with their evidential interpretation, the center of mass for the distribution of all p values shifts from large p values to small ones as the response probability increases. Differences are more easily discernible when considering their cumulative distribution functions (lower panel). Indeed all p values are valid (stochastically larger than the uniform distribution under the null) since their cumulative distribution functions (CDF) are smaller than that of the uniform distribution. This reflects the fact that all of the estimator-induced p values induce valid level-α tests. The inefficiency of the RBE estimator is reflected by the fact that the distribution of the corresponding p value under the alternative is far less peaked than for the other estimator-induced p values (flatter CDF). Although the PMEP-induced

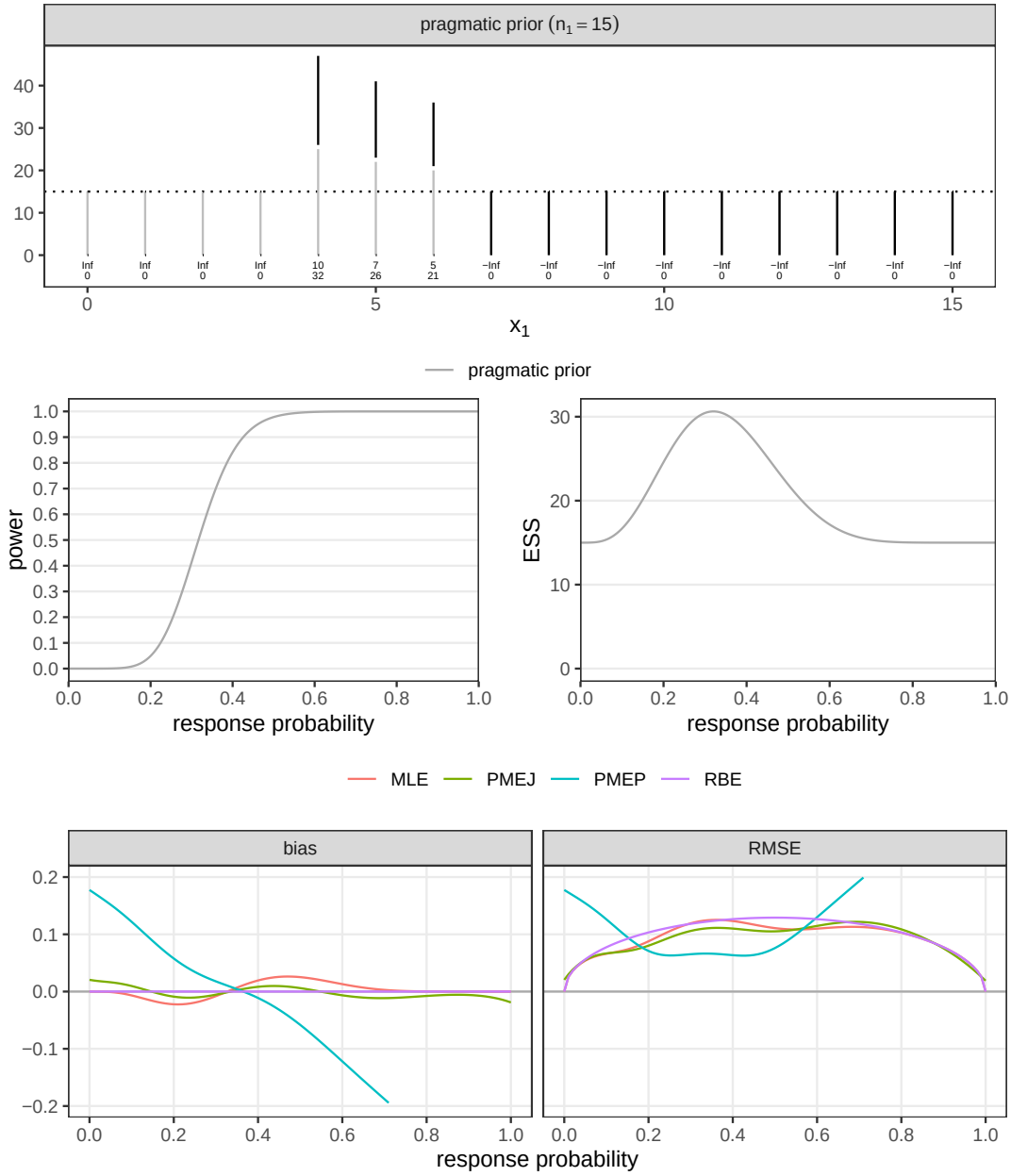

Figure 11.1.: Comparison of frequentist estimators (MLE and RBE) with posterior means for the pragmatic (PMEP) and Jeffreys prior (PMEJ) in terms of bias and root mean squared error (RMSE) for the design minimising expected sample size under the pragmatic prior and expected power constraint.

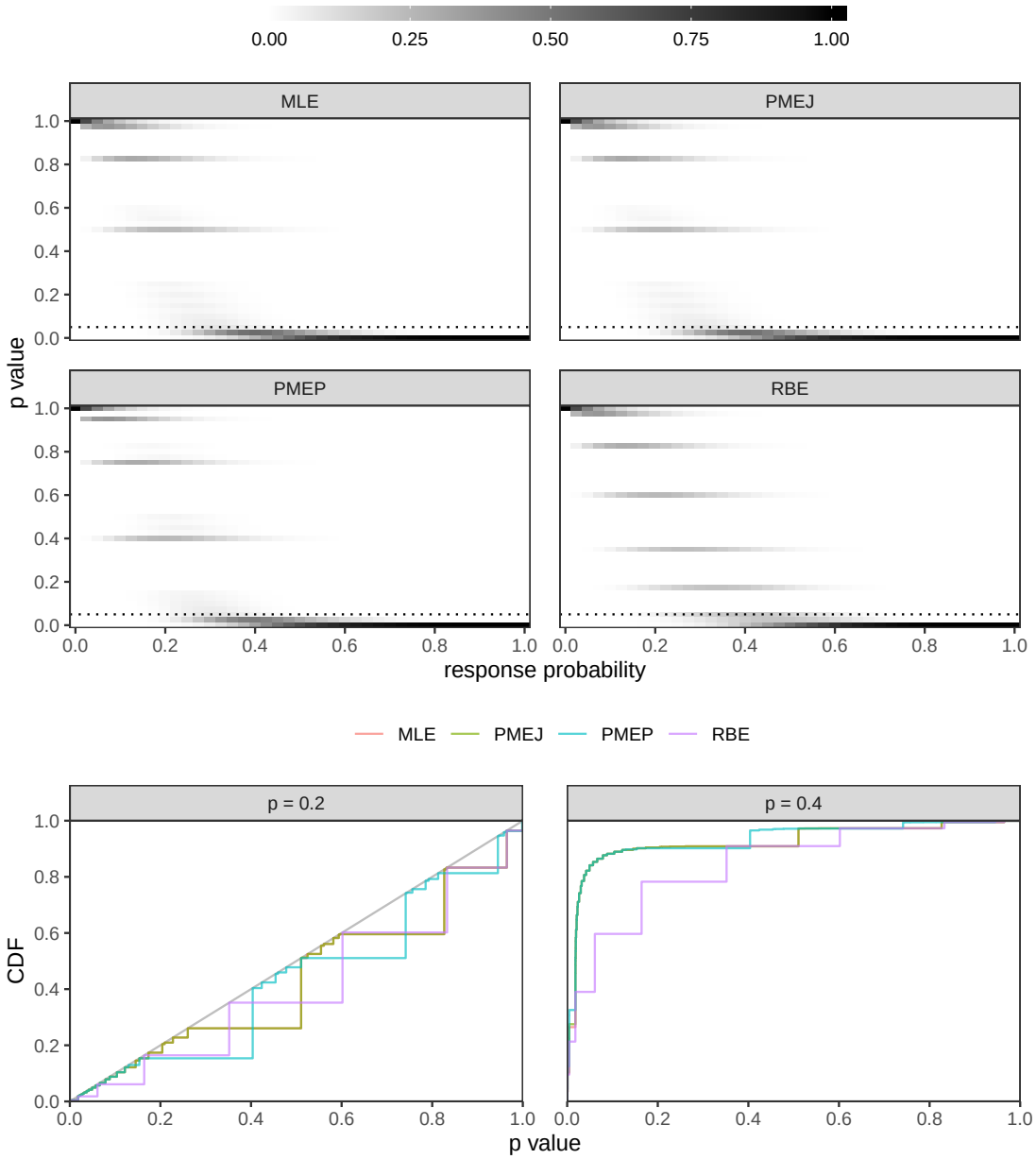

Figure 11.2.: Density plot of the estimator induced p values for the MLE, RBE, PMEJ, and PMEP point-estimators (top panel) and their distribution under the null (0.2) and a point alternative (0.4); under the null, the gray diagonal corresponds to the cumulative distribution function of the uniform distribution; the dotted line in the density plot panels corresponds to $\alpha = 0.05$; MLE and PMEJ overlap almost completely.

ordering also leads to a valid level-α test, it is clearly different from the MLE and PMEJ ordering since the corresponding CDF is different (most evidently so under the null).

Finally, Figure 11.3 compares the exact Clopper-Pearson type confidence intervals induced by the frequentist estimator (MLE and RBE) with the posterior credible intervals corresponding to the priors used for the Bayesian PMEJ and PMEP point estimators. Besides overall coverage, coverage for the upper and lower boundaries of the respective intervals are reported separately. Not surprisingly, the Bayesian credible intervals do not achieve the nominal coverage levels for all response probabilities. Also, the coverage probabilities for all interval estimators vary substantially due to the discreteness of the problem. Coverage for the pragmatic prior credible posterior interval and moderate response probabilities is only slightly undershooting the nominal level but quickly approaches 0 towards more extreme response probabilities since these are highly unlikely under the prior. Both frequentist intervals have, by definition, exact coverage. The lower panel of Figure 11.3 shows the corresponding mean width of the intervals as a function of the response probability. The RBE-induced confidence interval is the widest for most response probabilities of interest since the estimator only makes use of stage-one data. Both Bayesian intervals are much narrower which is the flipside of being allowed to undershoot the nominal coverage level. Interestingly, the pragmatic prior leads to the most uniform average width curve. This is due to the fact that the low *a priori* likelihood of extreme probabilities leads to a prior data conflict and thus higher posterior uncertainty than under the non-informative Jeffreys prior which puts substantial weight on extreme response probabilities.

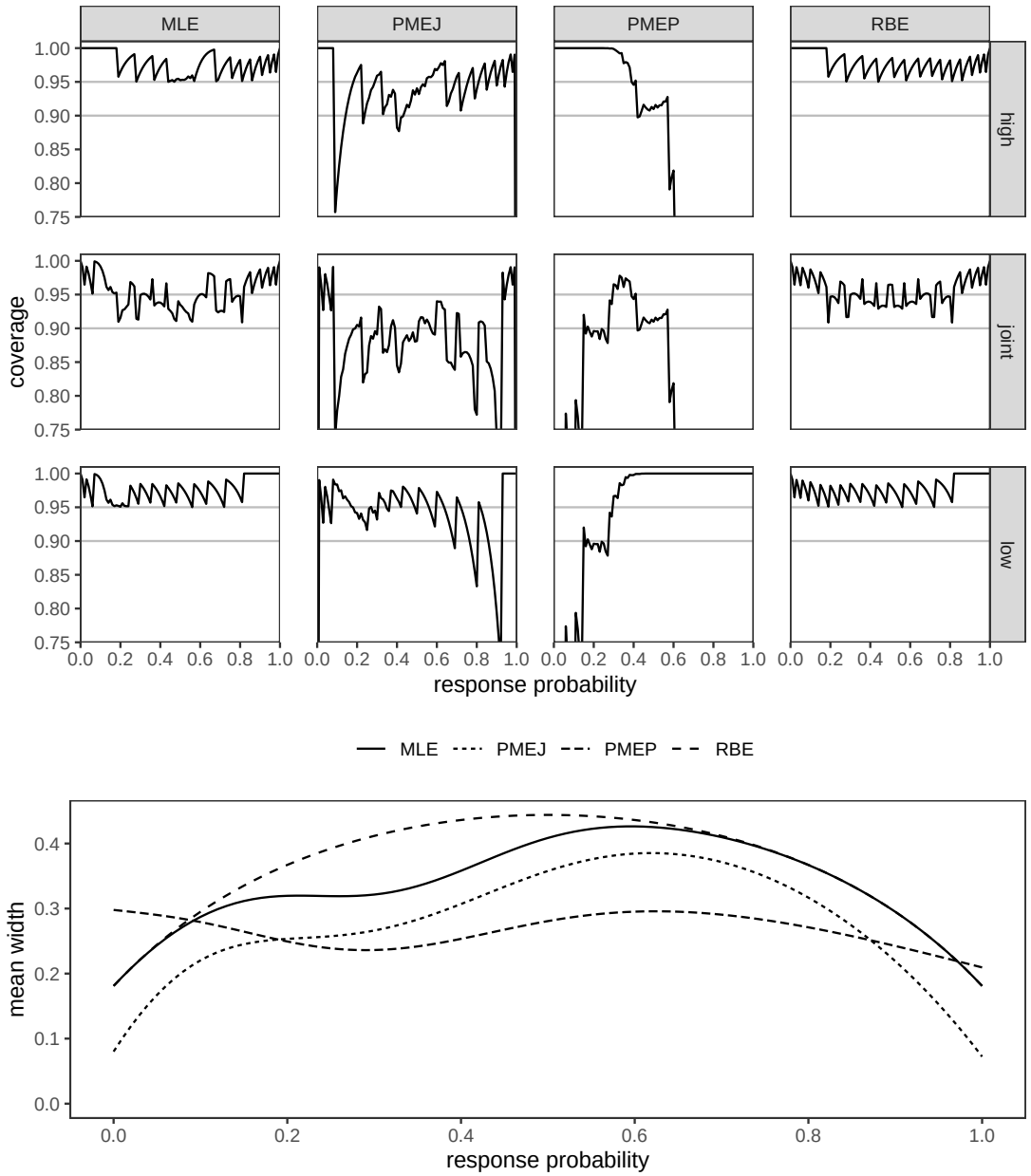

Figure 11.3.: Upper (high), lower (low), and joint coverage for Bayesian and frequentist interval estimators (top panel) as well as their mean width (bottom panel). For the Bayesian posterior mean estimators PMEJ and PMEP, the 90% credible intervals are given and 90% Clopper-Pearson type confidence intervals for the frequentist MLE and RBE.

11.2. Design incompatible p values

To illustrate the issues around test compatibility, consider a design proposed by Shan *et al.* (2016). They proposed to address the fact that sample size functions tend to be increasing in x_1 on the continuation region for designs optimising expected sample size under p_0 (see Section 2.3.1) by imposing a shape constraint on $n_2(\cdot)$ that forces the sample size function to be monotonically decreasing on the continuation region. Clearly, this approach contradicts the objective criterion and leads to rather peculiar designs that tend to be almost constant on the continuation region (i.e. close to group-sequential) in most situations. A particularly interesting design is obtained by applying the approach of Shan *et al.* to $p_0 = 0.6, p_{\text{alt}} = 0.8, \alpha = 0.05$ and $\beta = 0.1$. Their resulting optimal design (Shan) and the performance in terms of bias and RMSE of the MLE, the RBE, and the PMEJ for this design are compared in Figure 11.4. The characteristics of the estimators are similar to the results shown in Figure 11.1 considering the shift of the boundary of the null hypothesis from $p_0 = 0.2$ to $p_0 = 0.6$. For the Shan-design, the RBE does not reduce to the stage-one maximum likelihood estimator since the stage-two sample size function is not injective on the continuation region. The differences are, however, negligible.

More interestingly, the Shan-design is not compatible with the ordering induced by the MLE. To see this, Figure 11.5 shows the sample space of the design in x_1-x_2 coordinates together with the respective estimator-induced p values. Crosses indicated rejections under the estimator-induced test where the optimal design does not reject the null and, *vice versa*, circles non-rejection based on the induced p values where the design rejects the null. The degree of non-compatibility (number of potential outcomes that lead to a different test decision under the design and the estimator-induced test) is minimal for the MLE. In fact, only a single potential observation would lead to conflicting inference. By construction, the RBE is highly incompatible with the given design. Taking into account its *sub par* precision (see Figure 11.1) the RBE can thus not be recommended for general use with single-arm two-stage designs for binary endpoints. Interestingly, the PMEJ *is* compatible with the Shan-design in this situation.

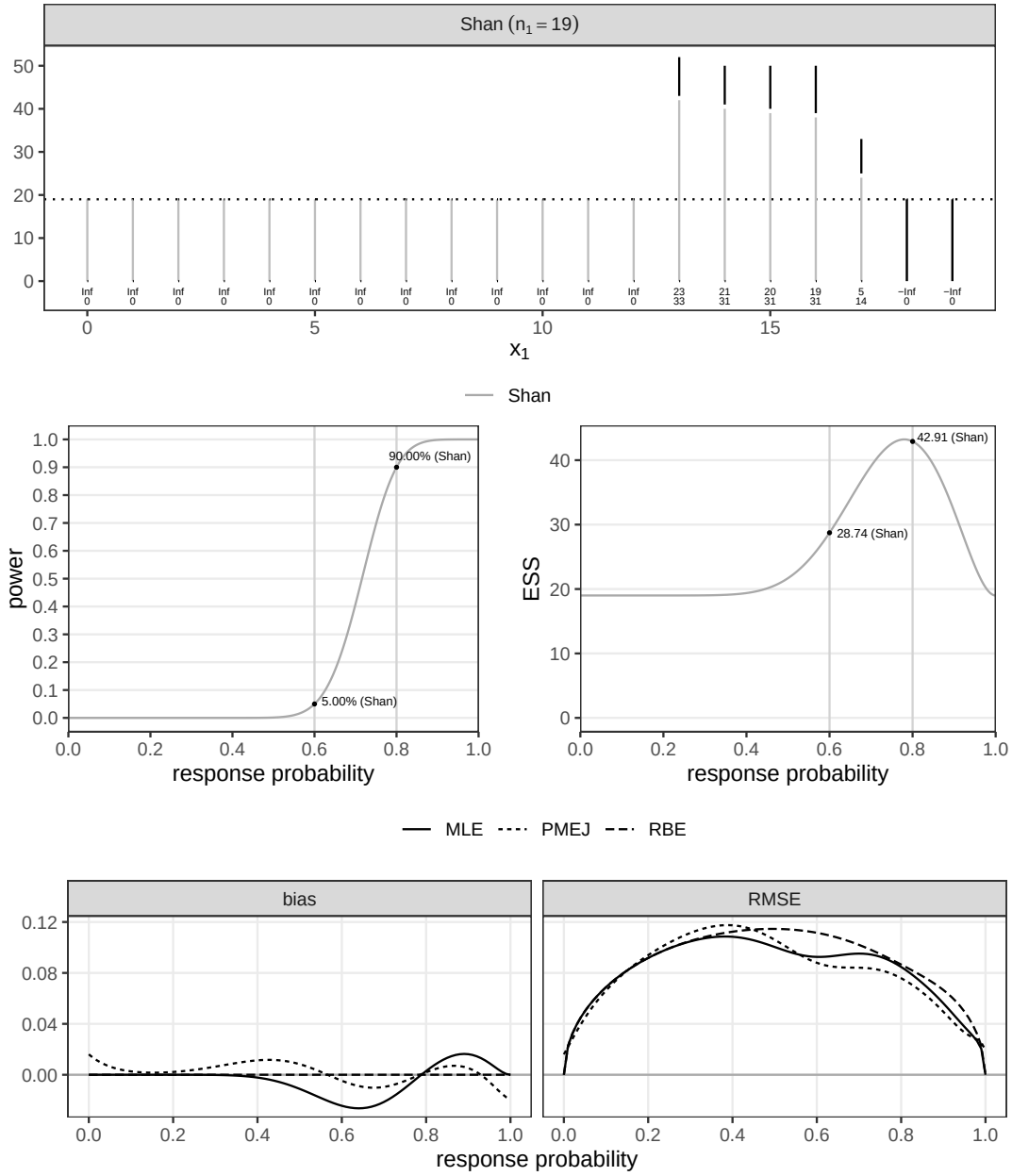

Figure 11.4.: Shan-design for $p_0 = 0.6, p_{alt} = 0.8, \alpha = 0.05$ and $\beta = 0.1$ and performance characteristics of the MLE, RBE, and the posterior mean estimator under the corresponding Jeffreys prior (PMEJ).

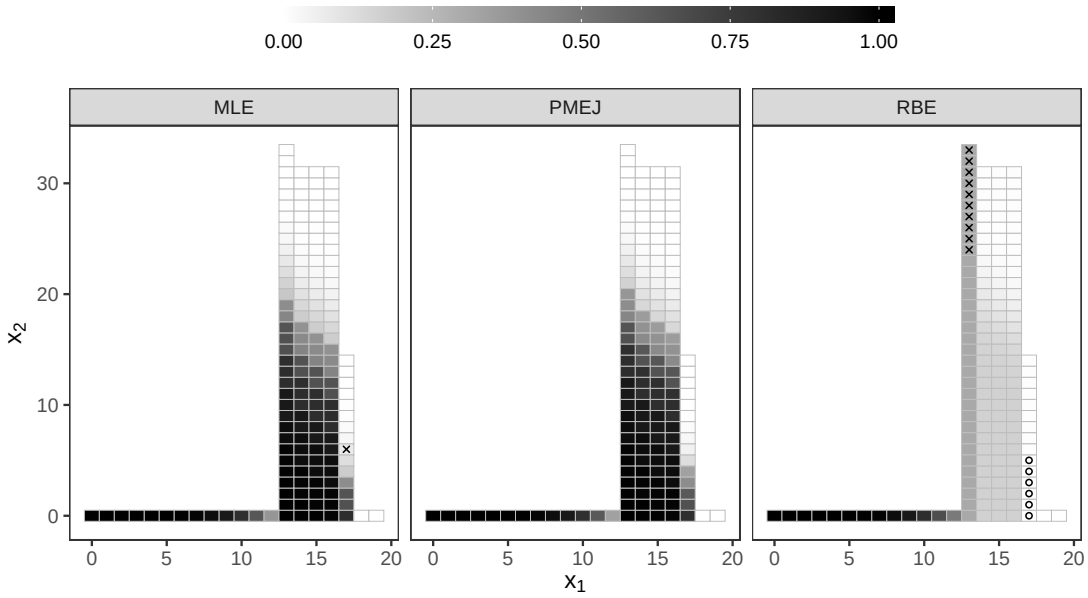

Figure 11.5.: Sample space of the Shan design (see Figure 11.4) and corresponding estimator-induced p values for the MLE, RBE, anf PMEJ point estimators; crosses indicated rejections under the estimator-induced test where the optimal design does not reject the null and, *vice versa*, circles non-rejection based on the induced p values where the design rejects the null.

11.3. Compatible maximum likelihood estimator

Incompatibility with the MLE occurs rarely when considering optimal design obtained from minimising expected sample size. For instance, the optimal designs minimising expected sample size under the alternative for $p_0 = 0.1, \ldots, 0.7$ and $p_{\text{alt}} = p_0 + 0.2$ with $\alpha = 0.05$ and $\beta = 0.2$ are all compatible with the MLE. When minimising expected sample size under the null instead, for $p_0 = 0.7$ the optimal design is again incompatible with the MLE for a single observation. Since the number of outcomes that lead to conflicting inference based on p values and the critical value function of the design is small even in cases where there are incompatibilities (e.g. the Shan-design discussed earlier, see Figure 11.5), the chances of actually encountering a situation where incompatibility might become a problem are thus small.

Coming back to the Shan-design, the compatible maximum likelihood estimator (CMLE) introduced in Section 5.4 can be computed. A direct comparison with the standard MLE is given in Figure 11.6. Since the required modification to make the MLE compatible with the design are extremely small, the overall performance characteristics of the CMLE are the same as those of the MLE (see Figure 11.4).

In terms of practical application, it is important to bear in mind that incompatibilities with the design's test decision can potentially occur when reporting p values for two-stage designs. These can be avoided by using a compatible estimator like the CMLE although the additional complexity of the procedure must be taken into consideration.

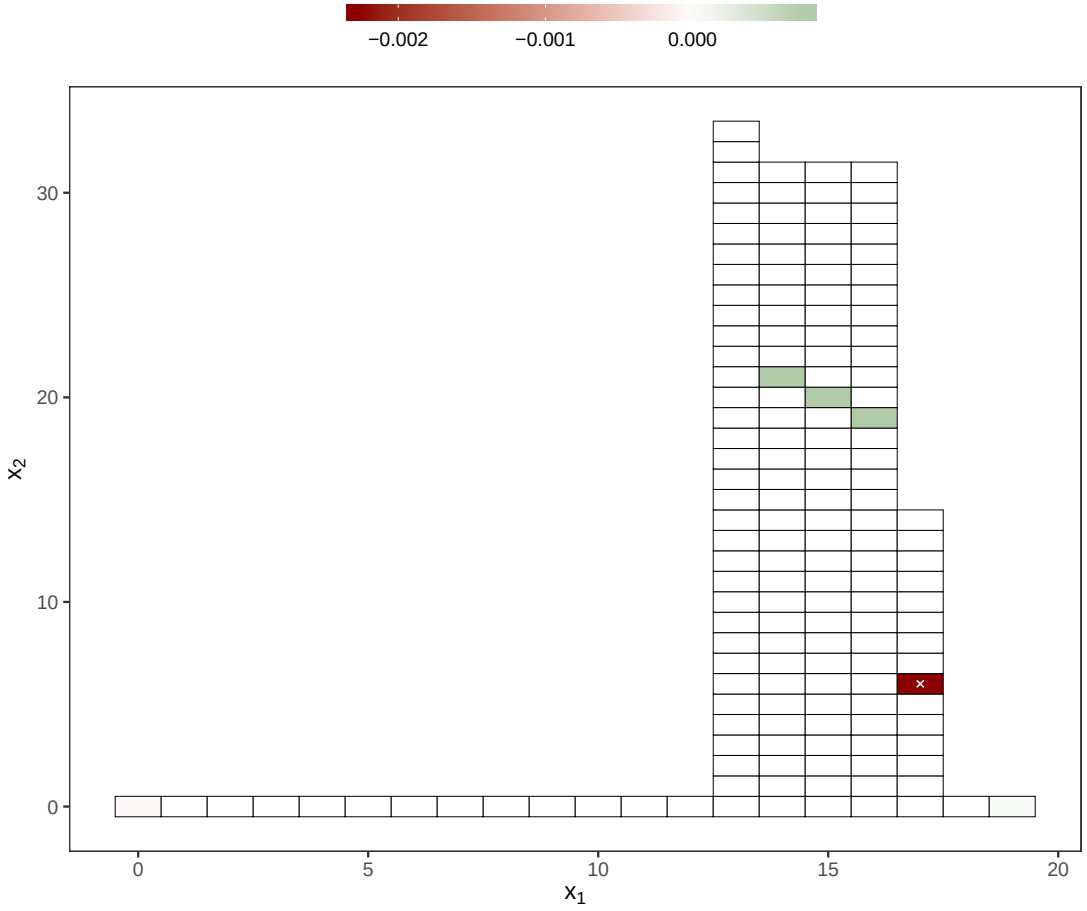

Figure 11.6.: Difference of the MLE and the CMLE on a per-observation basis; green corresponds to an increased estimate for the CMLE and red to a decrease as compared to the standard MLE; the white dot marks the outcome for which the MLE-induced p value is incompatible with the design decision.

12. Examples: Unplanned Adaptations

12.1. Unplanned adaptation in stage two

Consider the optimal design obtained as solution of the problem

$$\underset{n_1,\, n_2(\cdot),\, c_2(\cdot)}{\operatorname{argmin}} : \mathrm{E}_{\varphi(\cdot)}[\, n(X_1)\,] \tag{12.1}$$

$$\text{subject to :} \quad \Pr_{0.2}[\, X_2 > c_2(X_1)\,] \leq 0.05 \tag{12.2}$$

$$\Pr_{\varphi(\cdot)}[\, X_2 > c_2(X_1) \,|\, p \geq 0.3\,] \geq 0.8 \tag{12.3}$$

for the practical prior $\varphi = \varphi_{11.61,20.14,0.2|\leq 0.7}$ discussed in Section 9.1. Here $\alpha = 0.05$, $\beta = 0.2$, $p_0 = 0.2$, and $p_{\text{MCR}} = 0.3$. The resulting design is shown in Figure 12.1. The stage-one sample size of the design is 19 and $n_2(6) = 20$. Now assume that 6 responses out of 19 individuals are observed in stage one and another 2 out of 5 in stage two before a competing dose-finding study with less promising results is published. Assume that only 2 out of 10 individuals in the new study had a response. One way of incorporating this new information is by updating the informative component of the original pragmatic prior leading to $\varphi' = \varphi_{13.61,28.14,0.2|\leq 0.7}$. To conduct a sample size re-calculation as suggested above, the maximal conditional type one error rate and conditional expected power of the original design for this situation need to be determined. These are 0.164 and 0.698, respectively. Note that the favourable trial-internal results ($8/24 = 1/3 > 0.2$) before the recalculation lead to a larger conditional error than the unconditional $\alpha = 5\%$ level. Still, the results provide only little evidence that the effect is indeed relevant ($p_{\text{MCR}} = 0.3$) and the conditional expected power is thus less than the original threshold of 80%.

Since $\mathrm{E}_{\varphi(\cdot)}[\, n_1' + n_2'(X_1') \mid X_1' = 6, X_{(n_1', n_1'+5]} = 2] = n_1 + n_2'(6)$, one simply needs to find the smallest $n_2'(6)$ and the corresponding $c_2'(6)$ for which the conditional type one error rate is lower and the conditional expected power is larger than under the original design. In the example at hand, an exhaustive search yields $n_2'(6) = 24$ (instead of 20) and $c_2'(6) = 7$ (instead of 6) resulting in a new conditional type one error rate of 0.163 and a new conditional expected power of 0.727. Had the recalculation instead used the original $1 - \beta = 0.8$ as threshold for the conditional expected power, the recalculated stage-two sample size would have been 36 and the stage-two critical value would have changed to 10. In this case, the new second stage would have had a conditional error of 0.151 and a conditional expected power of 0.821.

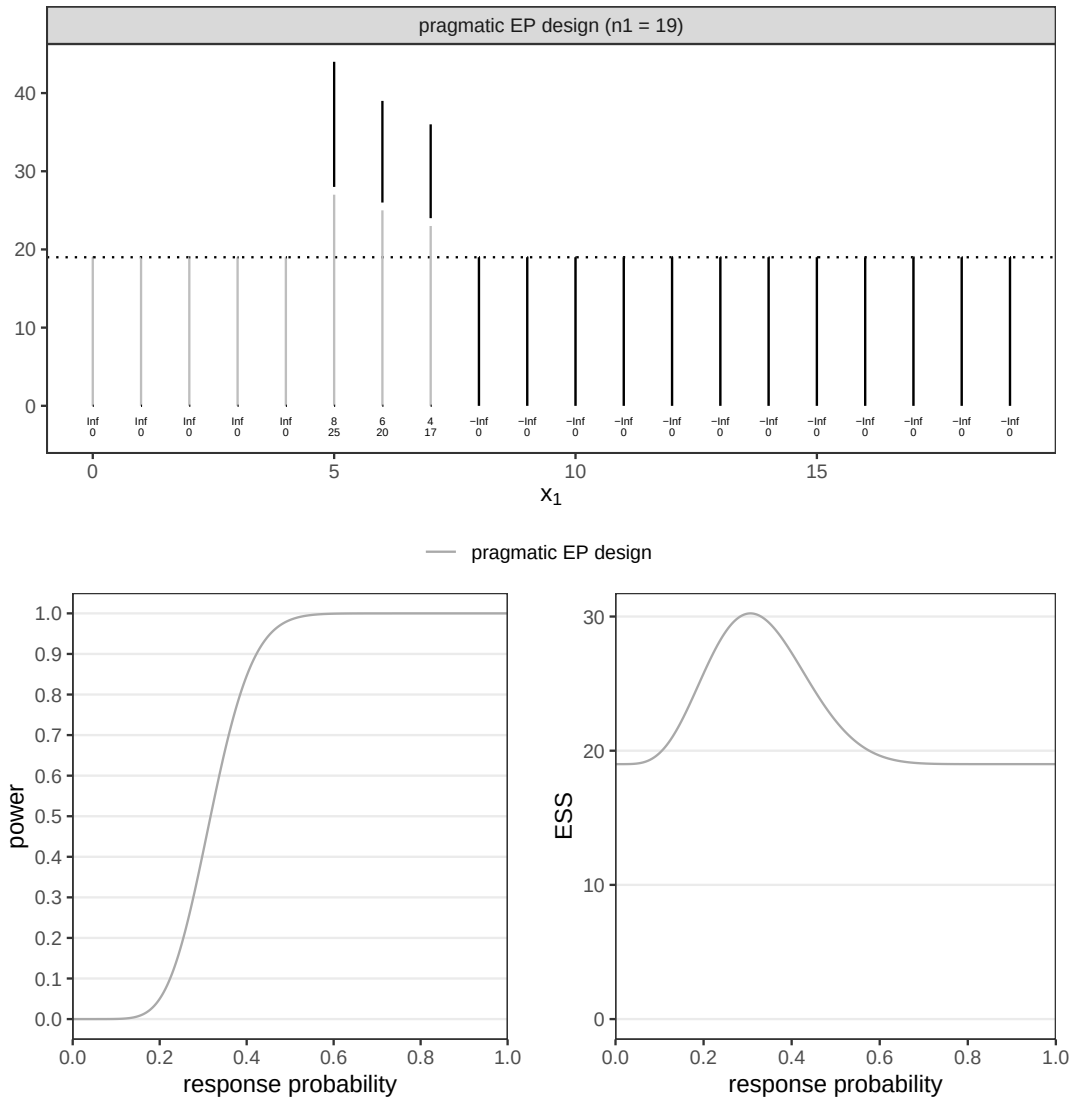

Figure 12.1.: Optimal two-stage design minimising expected sample size under the pragmatic prior and expected power constraint derived in Section 3.3.

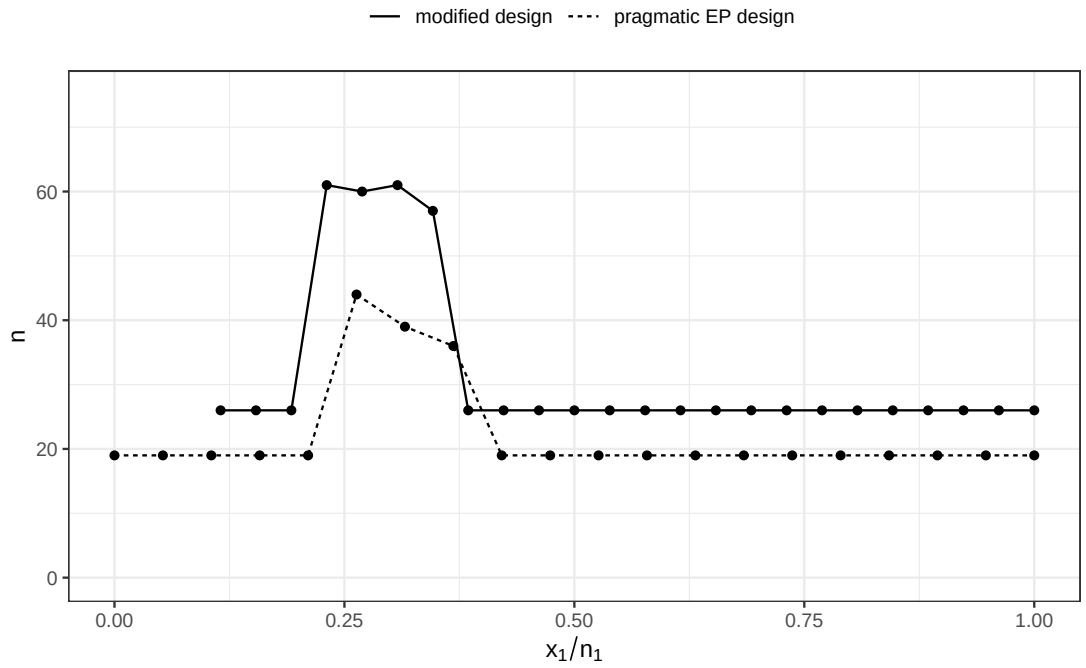

Figure 12.2.: Sample size function of the original design (pragmatic prior and expected power constraint, see Figure 12.1) and adapted partial sample size function after observing 3 out of 7 responses in stage one.

12.2. Unplanned adaptation in stage one

Using the same original design as in Section 12.1 one may consider a situation in which the additional external information is available earlier. Assume again that the prior is changed to $\varphi' = \varphi_{13.61,28.14,0.2}\big|_{\leq 0.7}$ but that only 3 responses out of the first 7 individuals are observed at the time point of adaptation. This means that the result of the adaptation must be a partial design with $x_1' = 3, 4, \ldots, n_1'$ and $n_1' \geq 7$. Solving problem (6.15)-(6.17) in this situation results in the partial design depicted in Figure 12.2. Instead of showing the original design and the modified one side-by-side only the (partial) sample size functions are plotted as line graphs to better capture the characteristic change. Since $n_1' = 26 \neq n_1 = 19$, the x-axis is scaled to the observed effect x_1/n_1 and x_1'/n_1' respectively. The fact that the modified design is only defined for $x_1' \geq 3$ is reflected in the sample size function of the adapted partial design only being defined for response rates greater than $3/26 \approx 0.12$. Since the new prior is more conservative (proportion of favourable results in pilot data is lower), the sample size function tends to increase for the modified design to maintain sufficient power.

13. Examples: The Continuous Case

Consider the randomised two-arm case of a superiority test for a difference in tumour volume after 3 months of intervention. Assume that $\theta_0 = 0$ (no difference) and that the minimal clinically relevant θ is a standardised difference of 0.2. Further assume that a power of 80% at $\theta = 0.2$ is required, and that the maximal type one error rate at $\theta = 0$ is 5%.

To illustrate the effect of the choice of objective criterion on the resulting optimal designs in the continuous case, three criteria are compared. Firstly, the expected sample size under the point alternative of $\theta = 0.2$ is minimised (alternative). Secondly, the design minimising expected sample size under the null (null) is considered. Thirdly, a design minimising expected sample size under a Gaussian prior with standard deviation of 0.1 and mean of 0.1 compromising between the null and alternative designs is constructed (Bayesian). Using the direct spline-based approach implemented in the `R` package `adoptr` (Kunzmann *et al.*, 2020b,c), the optimal two-stage designs were computed and are visualised in Figure 13.1. For two-stage designs with continuous test statistic, the sample size function $n(x_1)$ and the critical value function $c_2(x_1)$ can be depicted as line graphs without falling back to the slightly more involved representation chose for discrete designs in the remainder of this thesis.

The characteristic increasing shape of the sample size function when minimising expected sample size under the null hypothesis is the same as in the discrete case (see also Section 8.2). The shape is decreasing when minimising expected sample size under the alternative. Minimisation of expected sample size under the null also leads to an extremely high optimal efficacy boundary ($x_1 \approx 3.5$) as compared to the alternative and Bayesian designs. This, and the extremely large maximal sample size for the null design is due to the fact that the objective criterion is penalising sample size for large x_1 much less than under the other two criteria. Interestingly, the Bayesian criterion exhibits a unimodular sample size function - just as in the discrete example considered in Section 9.1 (see Figure 9.2). These differences in the sample size functions lead to very different expected sample size profiles as functions of θ (top right panel in Figure 13.1). The null-design only dominates the other two designs on and close to the null hypothesis at the expense of a massively increased sample size for large effect sizes. In comparison, the expected sample size profiles of the Bayesian and the alternative design are fairly similar with the Bayesian design exhibiting substantially lower sample size for small effect sizes at the cost of only minimally increased expected sample size for larger values of θ.

The power curves of all three designs are almost identical since the same constraint on the fixed alternative of $\theta = 0.2$ was used. The differences in sample size are thus only due to the different choice of objective criteria. Evidently, due to the large sample sizes, solving the continuous relaxation of the problem before rounding the

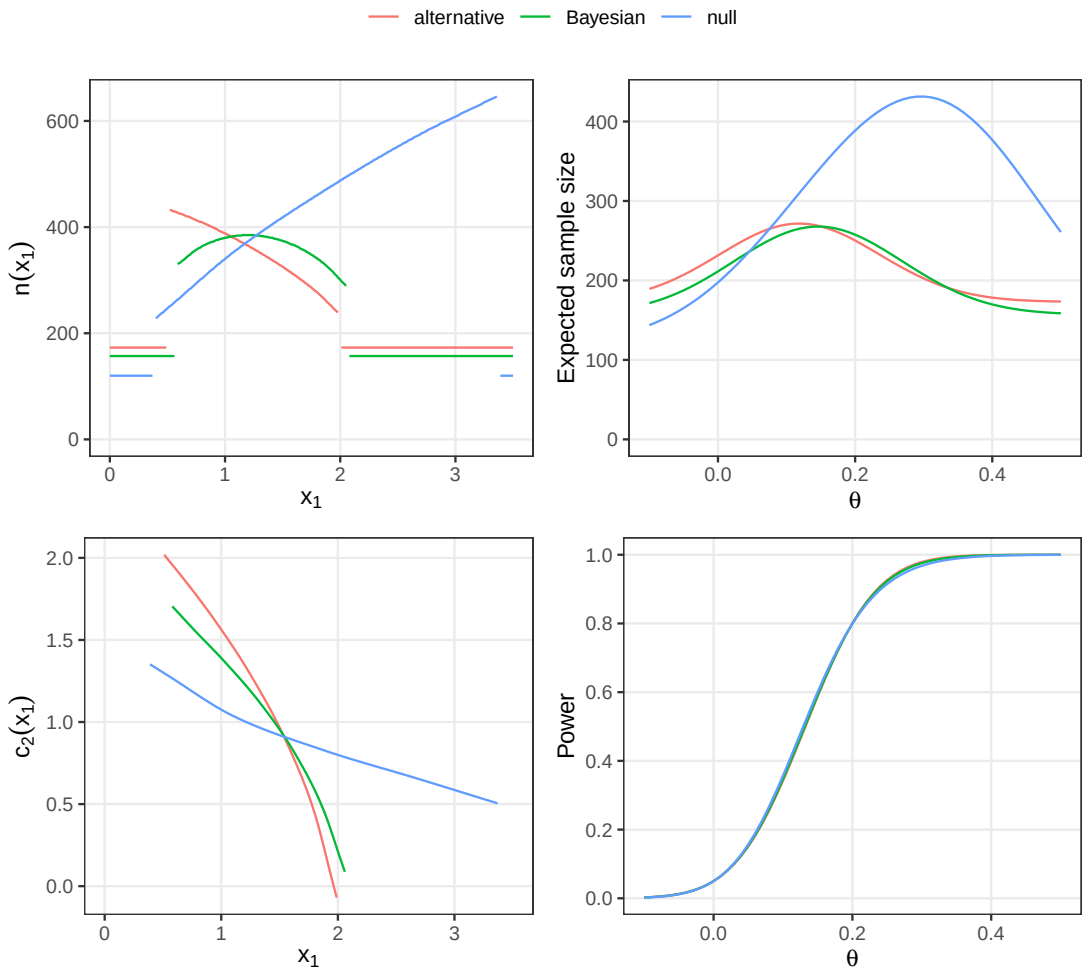

Figure 13.1.: Sample size function, critical value function, expected sample size curve, and power curve of the minimal expected sample size designs corresponding to the three different assumptions on θ discussed in Chapter 13.

sample sizes to the next integer values does not affect the power or type one error rate constraints negatively since both the type one and type two error rate constraints are met exactly (lower right panel in Figure 13.1).

This example again demonstrates that the choice of objective criterion is important and that the minimisation of expected sample size under only a single potential value of θ can lead to very different design characteristics. A design that optimises a weighted criterion in a Bayesian way is more robust under different values of θ.

Part IV.

Discussion and Summary

14. Discussion

14.1. Inadequacy of the single-stage binomial test

The starting point of this thesis was the inadequacy of the simple binomial test for single-arm clinical trials with binomial endpoint. Binary endpoints such as tumour response are common in early phase II trials in clinical oncology. Two properties make a simple binomial test ill-suited as design for a clinical trial with binary endpoint.

The first problem is of rather technical nature. Since the underlying test statistic is discrete, a binomial test may never fully exhaust a specified maximal type one error rate (α level). The significance level of a binomial test might thus be substantially lower than required (see Figure 1.1). If the permissible significance level (often 5% one-sided in phase II trials, (Simon, 1989; Ivanova *et al.*, 2016)) could be exhausted more fully, the required sample size for a given configuration could be reduced. From a purely statistical perspective, randomised tests could address this inefficiency by basing the final test decision on a biased coin toss if the trial result (number of responses) lies exactly on the critical value. The probability to still reject the null hypothesis even if the number of responses lies exactly on the decision boundary where a traditional binomial test could not reject the null hypothesis, can be chosen to match the specified α-level exactly. In clinical decision making, however, basing the results of an expensive trial on a (biased) coin toss is hardly acceptable.

The second problem is not specific to the standard binomial test but rather a general issue with trial designs that do not allow an interim assessment of the data from an initial stage. Clinical trials, and even more so early phase II trials, are planned under substantial uncertainty about the true effect size. Before the onset of a trial it is hard to judge whether planning assumptions hold and an interim assessment based on the first few data points is thus intuitively attractive. Such two-stage procedures have become common in clinical trials in general and early oncological phase II trials in particular (Jennison *et al.*, 2000; Ivanova *et al.*, 2016).

14.2. State of the art

Both shortcoming of the standard binomial test - its lack of efficiency and its lack of adaptivity to interim results - can be addressed by switching to a two-stage procedure. Early suggestions in the literature are group-sequential two-stage designs (Simon, 1989; Mander *et al.*, 2010). These designs allow the early termination of the ongoing trial at an interim analysis based on the number of observed responses up to this point under strict control of the overall type one error rate. Here, it is import-

ant to stress that the data-dependent interim decision could inflate the overall type one error rate if not properly accounted for during the planning stage (Bauer *et al.*, 2016). Group-sequential designs are characterised by a small number of parameters (four without, five with early stopping for efficacy). Still, these are already too many degrees of freedom to determine a design (i.e., its parameters) using only constraints on the desired type one and type two error rates as it is possible with a single-stage design that is determined by sample size and critical value alone. Instead, a quality criterion must be optimised subject to constraints on the type one and type two error rates to obtain a unique design. The choice of criterion is highly situational and different suggestions were put forward in the literature. Simon (1989) suggested to minimise the expected sample size on the boundary of the null hypothesis as a particularly attractive criterion in early oncological phase II trials since this corresponds to minimising the expected number of patients being treated ineffectively within the trial should the new treatment turn out to be less effective than treatment as usual. Mander *et al.* (2010), however, argued that this leads to unnecessarily long trials in cases where the new treatment shows a favourable response rate and proposed to instead minimise expected sample size under a point-alternative hypothesis.

While two-stage designs can also be seen as a form of randomised tests (see Section 1.3), the crucial difference to the single-stage biased-coin randomised test lies in the fact that the source of randomness is trial-internal: Once the interim outcome is observed, the final test decision is again a deterministic function of the stage two data. This is important to make such design acceptable to non-statisticians.

A natural extension to the concept of group-sequential designs is to allow the sample size and critical value to vary with the exact number of observed responses during the interim analysis instead of just determining whether to continue or to stop the trial. The corresponding optimisation problem is much harder to solve than the one for group-sequential designs since the number of parameters is substantially larger (see Englert *et al.* (2013) and Section 2.3). Previously, authors either solved the problem approximately (Banerjee *et al.*, 2006) or via custom implementation of the Branch & Bound algorithm for integer-value problems (Englert *et al.*, 2013; Shan *et al.*, 2016).

14.3. An effective solution method

The first problem addressed in this thesis is the reliable and efficient *exact* solution of problems arising from optimising generic two-stage single-arm designs for binary endpoints which allow the sample size and the critical value to depend on the exact number of observed responses at an interim analysis. The core ideas were put forward in (Kunzmann *et al.*, 2016) and substantially refined in Section 2.3. The proposed approach transforms the problem in a standard form for integer linear programs (ILP) which can then be solved using existing, highly-specialised solver software. While these solvers are still based on the Branch & Bound algorithm, they guide their transition of the search space by tried-and-tested heuristics and information about the problem structure obtained from the relaxation of the problem to real numbers. This allows much speedier solution and the extension of the search space not previously

possible with existing methods.

The novel approach also allows to consider additional 'nicety' constraints via a common formalism to ensure certain desirable properties of the solutions obtained. For instance, global unimodality of the sample size function can be incorporated as a constraint (see Section 2.3.1). While these constraints are not necessary *per se* they make it easier to communicate the resulting designs to practitioners and fix artefacts caused by the discreetness of the underlying test statistic.

Obtaining solutions relatively fast (global optima within seconds instead of minutes or hours) is important since it is necessary to evaluate and compare a set of designs under varying planning assumptions before reaching a final decision (see Section 8.1).

14.4. Optimal group-sequential designs

The marginal benefit of fully generic two-stage designs over group-sequential ones crucially depends on the scenario. Clearly, the minimax objective favours group-sequential solutions and for many situation with wide priors, the optimal two-stage design is not too variable (see Section 2.4). In practice, a group-sequential design will thus often be sufficiently effective and easier to obtain.

A major advantage of the generic two-stage design methodology rather lies in the fact that the corresponding optimal generic two-stage designs can reveal properties of a chosen objective function that would otherwise remain hidden when only considering group-sequential designs. A good example for this is the fact that the optimal sample size function when minimising expected sample size under the null hypothesis is generally increasing on the continuation region. This phenomenon was first noted by Banerjee *et al.* (2006) and is a direct consequence of the choice of objective criterion. Instead of criticising the properties of a corresponding optimal generic-two-stage design, the discussion should rather focus on the justification of the objective function itself and its properties.

With the effective solution method proposed in this thesis (see Section 2.3) the main argument in favour of using simpler group-sequential design is mostly refuted. Particularly in early clinical oncology, where the stakes for study participant are high, the most effective available solution (optimal generic two-stage design) should be used.

14.5. Optimisation under uncertainty

Optimal designs vary substantially depending on the assumptions on the unknown true response probability (see Section 8.2). Since the true response probability is unknown at the planing stage, a way of incorporating this uncertainty in the optimisation process is crucial. Instead of selecting a design that is only optimal for a particular response probability, the final design should rather balance between performance at different likely values of the response rate. Simon (1989) was already aware of this issue and proposed to either minimise the maximal sample size (worst case) of a design or to use a Bayesian objective criterion without going into details as to how

this should be done. Minimising the maximal sample size is an attractive option if no assumptions about the anticipated response probability can be justified. Otherwise, however, the resulting design may be less effective than a design that exploits prior knowledge about the anticipated response probability. Such prior knowledge can be integrated quantitatively into the planning process by adopting a Bayesian perspective. This requires the specification of a prior distribution for the unknown response probability which encodes the *a priori* relative likelihood of different response probabilities. Minimising expected sample size under such a prior distribution then reduces to minimising expected sample size under all possible response probabilities weighted by their respective *a priori* likelihood.

While this concept has been mentioned in the literature previously (Simon, 1989; Jennison *et al.*, 2015), its implementation for binary two-stage designs is another original contribution of this thesis. Chapter 3 explores the consequences of taking a Bayesian perspective to optimisation under uncertainty in great detail. The same Bayesian framework can be used to incorporate *a priori* uncertainty in the formulation of the power constraint. Concepts such as expected power (Brown *et al.*, 1987) or probability of success (Spiegelhalter *et al.*, 1986) to do so are discussed in the literature. Chapter 3 slightly generalises both definitions to better distinguish between the relevance and the *a priori* likelihood of response probabilities, explores their mathematical properties in detail, and highlights their close connection and their natural emergence in the context of utility optimisation. From the analysis in Chapter 3, it is clear that probability of success is ill-suited to be used as a replacement for traditional power constraints due to its dependence on the *a priori* probability of a relevant effect. Instead, expected power is a more direct extension of classical power constraint when uncertainty about the true response probability is to be reflected in the optimisation procedure. The concept of expected power allows a clear distinction between the minimal clinically relevant response probability and potential evidence in favour of larger response probabilities. In situations with prior evidence in favour of larger response probabilities, this approach then allows to formally justify smaller sample sizes than those required when powering on the minimal clinically relevant response rate.

The concept of probability of success, however, emerges naturally when considering utility maximising designs (see Section 3.4). Here, instead of specifying a target type one and type two error rate directly, a utility function is specified that links the potential outcomes of a trial (correctly rejecting the null, false positive finding) to corresponding rewards. The practical implementation of designs that directly minimise expected utility is hindered by the difficulty of defining a concrete utility function and the reward parameters. The utility-based approach can be reversed to uncover implicit assumptions about utility parameters. Such an approach can then be used to guide the choice of different standard power levels (80% or 90%?) by relating these back to implied real-world rewards. This additional information may be used to aid the decision making process for selecting the target power level of a design.

14.6. Optimisation *versus* unplanned recalculation

In the context of response-adaptive designs it is important to distinguish between (optimal) generic two-stage designs and methods for the unplanned recalculation of a trial's sample size. Both concepts are often mixed but quite different. For instance, Jennison *et al.* (2015) optimise a trial's design for continuous outcomes with respect to a utility score but use a combination function approach to control the type one error. This is clearly ineffective (Pilz *et al.*, 2019) and a direct optimisation of all design parameters - as discussed in the Chapters 2 and 3 - is superior. Even though the sample size and the critical value are response-adaptive in a generic two-stage design, the design itself is still completely pre-specified. Unplanned adaptations, on the other hand, are by definition not pre-specified but the methodology is often used to define binding adaptation rules *a priori* or parts of them (Jennison *et al.*, 2015) which is less effective than a direct optimisation. In fact, adaptations based solely on trial-internal data are never necessary within an optimal two-stage design (see Chapter 6) since an optimal design considers all potential stage-one outcomes and the corresponding ideal decision before the start of the trial. Unplanned adaptations do, however, play an important role in handling unforeseeable external changes or violations of the pre-specified sampling protocol. The methodology is still necessary to react to newly emerging external information that might invalidate the original planning assumptions.

A common issue with unplanned design adaptation is the choice of adaptation criterion (Bauer *et al.*, 2016). The approach to unplanned design adaptations for optimal designs taken in Chapter 6 resolves this ambiguity by deriving the adaptation criterion directly from the original objective criterion that was defined during the planning phase of the trial. The full flexibility of unplanned design adaptations is, of course, still available. This means that one could, in principle, also use a completely different objective criterion during the adaptation. Yet it seems natural to maintain the original objective criterion and just adjust it such that it reflects changes in study-external evidence more appropriately.

The conditional error methodology used in Chapter 6 to realise these unplanned adaptations is similar to the approach in (Englert *et al.*, 2015). Other than Englert *et al.* (2015), the result of the adaptation is not just a conditional error function that still needs to be mapped to actual sample sizes and decision boundaries, but an immediately usable sample size (function) and critical value (function) that are optimal under the original criterion conditional on the data observed at the time point of the adaptation and taking potentially new trial-external evidence into account.

The availability of optimal flexible adjustment methods is crucial to a more widespread use of effective optimal two-stage methods in clinical trials generally. Investigators will be much more likely to adopt novel, more effective methods if they can be sure that a trial is able to react to practical necessities or newly emerging evidence. This is particularly important in a competitive and fast-paced environment such as oncology. The main contribution of this thesis in this respect is to map existing methods (conditional error principle) to the particular situation and to develop an *optimal* way of conducting the necessary adaptions.

14.7. Inference

The main focus of this thesis are optimal two-stage designs for binary endpoints, guidance on the choice of objective functions, and the incorporation of uncertainty in both the objective and the necessary power constraint.

Still, while the design of a clinical trial is important to reach the study objectives effectively, so is proper inference after the trial is completed. In the context of two-stage designs, a complications for traditional frequentist inference is given by the fact that the test statistic is two-dimensional. This implies that there is no natural ordering in terms of extremeness from the null hypothesis for two-stage designs. An ordering, however, is crucial for the definition of a p value, one of the most widely used measures of evidence in clinical statistics. This problem also affects the Clopper-Pearson approach to confidence-intervals. Several more or less arbitrary choices for defining an ordering on the outcome space are discussed in the literature (Jennison *et al.*, 2000; Wassmer *et al.*, 2016). In Chapter 5 of this thesis, an argument for the estimator-based ordering is put forward. The main argument is, that it is the ordering that preserves compatibility properties between p values, test decision, point estimates, and confidence intervals such as the fact that the null hypothesis should only be rejected if the p value is smaller or equal to the chosen significance level or that the $(1 - 2\alpha)$-two-sided-confidence interval only intersects with the null hypothesis if the test does not reject the null hypothesis. To achieve this, the p values induced by an estimator via its natural ordering must be compatible with the test decision of the underlying design. This is impossible to achieve for mean-unbiased estimation but mostly the case for the commonly used maximum likelihood estimator (MLE). In cases where the MLE is (partially) incompatible with the underlying design, a minimally distorted version of the MLE is derived that gives rise to a design-compatible ordering. This slight modification then guarantees that p values are smaller or equal to the chosen significance level if and only if the design rejects the null hypothesis and that the corresponding Clopper-Pearson-type confidence interval overlaps with the null hypothesis if and only if the design rejects the null. These compatibility properties are by no means statistically necessary but they are often expected by practitioners. Hence, their violation may lead to confusion about the interpretation a trial's results. Being able to avoid such situations entirely is thus a major benefit.

A major problem with frequentist inference after *unplanned* design adaptations is the fact that the result of an unplanned adaptation is only a partial design conditional on the data observed so far. It is thus impossible to tell with certainty what one would have done, had the data at the time-point of adaptation been different. This is, however, necessary to derive valid p values that are compatible with the final test decision. Methods for obtaining valid p values by, e.g., a combination function approach can still be used but the advantages of a compatible inferential framework for p values, point estimate, confidence interval, and test decision are lost. The ideal way of conducting compatible frequentist inference after an unplanned design adaptation thus remains an open problem.

Alternatively, Chapter 4 explores how the Bayesian paradigm (Jeffreys, 1998) can be used to draw inference after the conclusion of the trial. This is particularly attractive if

the necessary prior was already specified during the planning phase (see Chapter 3). Under the Bayesian paradigm, the notion of compatibility is not meaningful since the full posterior distribution is reported instead of a single p value. The necessarily subjective planning prior (see Section 9.1) introduces bias in posterior mean point estimates (see Section 10.1). This can be avoided by using a different prior distribution for the analysis of the trial than the one used for its planning (O'Hagan *et al.*, 2005). Objective Bayes theory addresses this issue via 'non-informative' priors. In single-parameter problems such as the one at hand, the Jeffreys prior is often employed as analysis prior (Jeffreys, 1946). A detailed analysis of the properties of the design-specific Jeffreys prior in Section 10.1, however, shows that the relative likelihood of different response probabilities implied by the Jeffreys prior might sometimes be hard to justify.

A major advantage of the Bayesian paradigm is that inference is unaffected by unplanned changes to the sampling scheme. This is due to the fact that all Bayesian inference is solely based on the likelihood and the chosen prior - not the design itself. The problems surrounding the definition of an ordering to obtain a valid p value consequently do not apply when reporting posterior probabilities instead of p values. Note that the use of posterior probabilities instead of p values does not, in any way, affect type one error rate control since this is a property of the optimal design itself and independent from the inference following the trial.

Furthermore, Bayesian posterior-mean point estimates showed favourable properties in terms of their bias-variance trade-off (see Section 5.2). This is due to the fact that the posterior mean estimator directly minimises the expected finite-sample quadratic error, whereas the maximum likelihood estimator is mostly popular due to its asymptotic properties. For large sample sizes and a non-informative prior choice, the differences are negligible, though.

14.8. Conclusions

Optimisation of trial designs can lead to more effective studies in terms of length and costs (both directly related to sample size). In clinical trials, two-stage designs are certainly preferable. This is particularly the case for binary endpoints often used in early oncological studies since the one-stage binomial test is ineffective. Optimal generic two-stage designs are an attractive choice since they fully exploit the information available during an interim analysis as compared to group-sequential design which only use this information to determine whether or not to stop the trial early. Similarly important, it can be insightful to study the optimising designs for different objective criteria to get a better understanding for the properties of the objective criteria. With the technical challenges around computing optimal two-stage designs largely resolved, the discussion should shift towards the choice of objective criterion. Here, it is important to correctly incorporate planning uncertainty. Bayesian methods can be used to do so and balance between performance under different response probabilities. This is essential to obtain designs that are robust towards small deviations from planning assumptions while still exploiting available *a priori* information as effectively as possible.

Pre-specified optimal designs are always superior to *post hoc* adaptations using methodology for unplanned design adaptations. The latter, however, do play an important role in incorporating newly emerging trial-external information or to react to operational necessities.

Standard frequentist inference in two-stage designs is complicated by the choice of ordering on the outcome space. Whenever possible, a design-compatible point estimator and its corresponding induced p value and confidence interval should be used to avoid contradictions between p value and confidence interval on the one side and the actual test decision on the other side. Mean unbiased point estimation in generic-two stage designs with binary endpoint is possible but the variance of the estimator is high rendering it an unattractive choice if precision is of any concern.

Bayesian inference based on the posterior distribution of the planning prior is an attractive alternative since the planning prior will generally be published together with the study protocol. Although the prior remains subjective, the fact that it is pre-specified will increase the credibility of inference drawn from it. The fact that Bayesian inference is unaffected by unplanned design changes makes this a particularly attractive choice when there are reasons to expect the necessity of a design change during the course of the trial. Furthermore, posterior point estimators tend to exhibit favourable properties in terms of their trade-off between bias and variance.

Software implementations for all methods discussed in this thesis are made available and are discussed in more detail in Appendix A.1. The source code underlying all examples discussed in this thesis is available in an online repository and permanently accessible via a digital object identifier. For details on the reproducibility concept of this thesis, see Appendix A.2.

15. Summary

This thesis presents a novel, fast, and exact way of computing optimal two-stage designs for single-arm trials with binary endpoint. These designs are commonly used in early oncological trials to asses the tumour response rate under a new therapy. Here, a 'response to treatment' is usually defined in terms of the RECIST guideline. The new method allows solving problems of any practically relevant size in this field within seconds or minutes which is a substantial speed-up as compared to previous implementations. This makes it easier to explore different planning scenarios interactively.

Most clinical trials have to be planned under substantial uncertainty about the true effect size. This is particularly important in early-stage trials where little prior evidence is available. It is thus important to devise methods that can incorporate uncertainty into the planning process. This thesis develops and presents multiple ways to do so and thus extends the classical method for sample size calculation which assumes a fixed point alternative. The principled incorporation of uncertainty is achieved by using a Bayesian prior probability that weighs plausible response rates by their *a priori* likelihood. In general, larger uncertainty about the true response rate then leads to optimal designs with less variable stage-two sample size functions. This makes the resulting designs more robust towards mis-specification of the expected response rate.

A brief outlook as to how these ideas can be extended to continuous outcomes is given at the end of this thesis. Continuous outcomes in early oncological trials arise, e.g, when it is more appropriate to consider the actual tumour volume as a continuous endpoint instead of a binary 'response criterion' like RECIST.

The problem of post-trial inference is discussed from both Bayesian and frequentist perspectives. In the frequentist case, a major challenge lies in the definition of p values that are compatible with the optimal-design's test decision in the sense that $p \leq \alpha \Leftrightarrow$ Test rejects the null hypothesis. While it is well-known that p values for multistage designs are not uniquely defined, in this thesis, it is argued that a definition in terms of 'test-compatible' point estimators has great practical advantages. Only if p values are based on estimators with such a property, the relations between point estimators, p values, and confidence intervals known from single-stage designs are preserved. For most sensible designs, the classical maximum likelihood estimator is test-compatible but not for all. A simple criterion for checking test-compatibility of an estimator during the planning phase and a general way of obtaining compatible estimators is described. It is further argued that unbiased estimation has unfavourable properties in the setting of binary two-stage designs and should be avoided.

Finally, this thesis also covers methods to deal with the need to react to unanticipated changes of the planning assumptions. In principle, optimal two-stage designs

need not be adjusted during the course of a study. A need to modify the design of an ongoing trial might arise if new trial-external information becomes available or the planning assumptions change otherwise. Here, it is important to distinguish unplanned design adaptations from the fact that the sample size itself *is* adaptive but in a pre-specified way, i.e. the design itself remains unchanged.

The methods outlined in this thesis therefore cover the entire life-cycle of a single-arm design with binary endpoint which are commonly used in early oncological trials. Software implementations of all methods discussed are provided (see Appendix A.1) and all examples presented in this thesis are available as interactive notebooks online (see Appendix A.2).

An interactive web app for exploring optimal designs in a few common settings without requiring any programming skills.

16. Zusammenfassung

Diese Arbeit stellt eine neuartige, schnelle und exakte Art der Berechnung von optimalen zweistufigen Designs für einarmige Studien mit binärem Endpunkt vor. Solche Designs werden häufig in frühen onkologischen Studien verwendet. In diesen Studien wird üblicherweise die Rate untersucht mit der eine bestimmte Tumorart auf eine neuartige Therapie in der Zielgruppe anspricht (die sogenannte 'response rate' oder Erfolgsrate). Der Therapieerfolg wird dabei häufig gemäß der RECIST Richtlinien bestimmt. Diese neue Methode zur Berechnung solcher optimalen Designs erlaubt die Lösung von Problemen für jede praktisch relevante Studiengröße innerhalb kürzester Zeit. Dies bedeutet eine erhebliche Beschleunigung im Vergleich zu früheren Implementierungen und erleichtert somit die interaktive Untersuchung verschiedener Planungsszenarien.

Klinischen Studien müssen üblicherweise unter erheblicher Unsicherheit bezüglich der relevanten Parameter (hier die tatsächliche Erfolgsrate) geplant werden. Dies trifft insbesondere auf Studien in frühen klinischen Phasen eines Entwicklungsprogramms zu, da hier in der Regel zum Planungszeitpunkt kaum datenbasierte Evidenz über die Erfolgsrate vorliegt. Methoden, die diese Planungsunsicherheit berücksichtigen können, sind daher von besonderem Interesse. Die vorliegende Arbeit entwickelt und präsentiert verschiedene Möglichkeiten solche Unsicherheiten in die Studienplanung zu integrieren und erweitert damit die klassische Methode zur Berechnung der benötigten Fallzahl, die üblicherweise von einer festen Punktalternative ausgeht. Zu diesem Zweck wurde ein Bayes'scher Ansatz gewählt, der plausible Erfolgsraten gemäß ihrer Vorabwahrscheinlichkeit gewichtet. Im Allgemeinen führt dabei größere *a priori* Unsicherheit zu optimalen Designs mit weniger variablen Fallzahlen für die zweite Stufe des Designs. Die resultierenden Designs sind also robuster gegenüber Fehlspezifikationen der erwarteten Erfolgsrate.

Die in dieser Arbeit entwickelten Konzepte lassen sich von der Situation der binären Endpunkte auf kontinuierliche Endpunkte übertragen. Obgleich dies nicht zentraler Bestandteil der vorliegenden Arbeit ist, wird die Vorgehensweise kurz anhand eines Beispiels erläutert. Kontinuierliche Endpunkte können in frühen onkologischen Studien zum Beispiel dann von Interesse sein, wenn anstelle eines binären Erfolgskriteriums gemäß der RECIST Richtlinien das tatsächliche Tumorvolumen (oder dessen Entwicklung im Zeitverlauf) betrachtet werden soll.

Neben der optimalen Planung einer Studie ist die effektive Auswertung nach Abschluss der Datenerhebung entscheidend für den Studienerfolg. Die finale statistische Inferenz wird sowohl aus der Bayesianischen als auch aus der frequentistische Perspektive betrachtet. Im Fall der frequentistischen Betrachtungsweise besteht eine wesentliche Herausforderung in der Definition von p Werten, die mit der Testentscheidung des zugrundeliegenden optimalen Designs in dem Sinne kompatibel sind, dass

$p \leq \alpha \Leftrightarrow$ Test lehnt die Nullhypothese ab. In der vorliegenden Arbeit wird argumentiert, dass eine Definition des p Wertes in Bezug auf 'Test-kompatible' Punktschätzer praktische Vorteile hat. Nur wenn p Werte auf Schätzern mit einer solchen Eigenschaft basieren, bleiben die Beziehungen zwischen Punktschätzern, p Werten, und Konfidenzintervallen erhalten, die aus einstufigen Tests bekannt sind. Für die meisten praktisch sinnvollen Designs ist der klassische Maximum-Likelihood-Schätzer Test-kompatibel, allerdings nicht in allen Situationen. Die vorliegende Arbeit zeigt, wie Test-Kompatibilität eines Schätzers während der Planungsphase überprüft werden kann und beschreibt einen allgemeinen Weg, um kompatible Schätzer zu erhalten. Weiterhin wird argumentiert, dass unverzerrte Schätzer in zweistufigen Designs in der Regel ineffizient sind und nicht zu Test-kompatiblen p Werten führen.

Schließlich betrachtet diese Arbeit auch Methoden, um auf unvorhergesehene Änderungen der Planungsannahmen zu reagieren. Prinzipiell sollten optimale zweistufige Designs zwar nicht während einer laufenden Studie modifiziert werden, da sie die Fallzahl bereits optimal an die beobachteten Zwischenergebnisse anpassen. Eine Designanpassung kann jedoch nötig werden, falls etwa neue Studien-externe Informationen verfügbar werden, die den Planungsprior obsolet machen. Hierbei ist es wichtig nicht-vorgeplante Designanpassungen von der vorgeplanten Anpassung der Stichprobengröße zu unterscheiden. Bei letzterer bleibt das Design selbst unverändert.

Die in der vorliegenden Arbeit beschriebenen Methoden decken also alle aus statistischer Sicht relevanten Projektphasen einer frühen einarmiges Studie mit binärem Endpunkt ab. Software-Implementierungen aller diskutierten Methoden stehen zur Verfügung (siehe dazu Anhang A.1). Alle in dieser Arbeit vorgestellten Beispiele sind als interaktive 'Notebooks' online verfügbar (siehe Anhang A.2).

Eine interaktive Web-App erlaubt das Experimentieren mit optimalen Designs, ohne jegliche Programmierkenntnisse.

17. Bibliography

Banerjee, A., Tsiatis, A. A. (2006). **Adaptive two-stage designs in phase II clinical trials**. Stat Med, *25*(19), 3382–3395.

Battelle Technology Partnership Practice (2015). **Biopharmaceutical industry-sponsored clinical trials: impact on state economies**. Technical report, Battelle Technology Partnership Practice. [accessed 2019-06-10].
URL `http://phrma-docs.phrma.org/sites/default/files/pdf/biopharmaceutical-industry-sponsored-clinical-trials-impact-on-state-economies.pdf`

Bauer, P. (1989). **Multistage testing with adaptive designs**. Biometrie und Informatik in Medizin und Biologie, *20*, 130–148.

Bauer, P., Bretz, F., Dragalin, V., König, F., Wassmer, G. (2016). **Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls**. Stat Med, *35*(3), 325–347.

Bauer, P., Kieser, M. (1999). **Combining different phases in the development of medical treatments within a single trial.** Stat Med, *18*(14), 1833–1848.

Bauer, P., Köhne, K. (1994). **Evaluation of experiments with adaptive interim analyses**. Biometrics, *50*(4), 1029–1041.

Berger, J. O., Bernardo, J. M., Sun, D. (2009). **The formal definition of reference priors**. Ann Stat, *37*(2), 905–938.

Bezanson, J., Edelman, A., Karpinski, S., Shah, V. B. (2017). **Julia: A fresh approach to numerical computing**. SIAM review, *59*(1), 65–98.

Brannath, W., Gutjahr, G., Bauer, P. (2012). **Probabilistic foundation of confirmatory adaptive designs**. J Am Stat Assoc, *107*(498), 824–832.

Brown, B. W., Herson, J., Atkinson, E. N., Rozell, M. E. (1987). **Projection from previous studies: a Bayesian and frequentist compromise**. Control Clin Trials, *8*(1), 29–44.

Byrd, R. H., Nocedal, J., Waltz, R. A. (2006). **Knitro: An integrated package for nonlinear optimization**. In **Large-scale nonlinear optimization**, pp. 35–59. Springer, Boston, MA.

Chang, M. N., Therneau, T. M., Wieand, H. S., Cha, S. S. (1987). **Designs for group sequential phase II clinical trials**. Biometrics, *43*(4), 865.

Clopper, C. J., Pearson, E. S. (1934). **The use of confidence or fiducial limits illustrated by in te case of the binomial**. Biometrika, *26*(4), 404–413.

COIN-OR initiative (2020). **CBC**. doi:10.5281/zenodo.3246628.

Cook, T. D. (2002). **P-value adjustment in sequential clinical trials**. Biometrics, *58*(4), 1005–1011.

DasGupta, A., Cai, T. T., Brown, L. D. (2001). **Interval Estimation for a Binomial Proportion**. Stat Sci, *16*(2), 101–133.

Dunning, I., Huchette, J., Lubin, M. (2017). **JuMP: a modeling language for mathematical optimization**. SIAM Review, *59*(2), 295–320.

Englert, S., Kieser, M. (2013). **Optimal adaptive two-stage designs for phase II cancer clinical trials**. Biom J, *55*(6), 955–968.

Englert, S., Kieser, M. (2015). **Methods for proper handling of overrunning and underrunning in phase II designs for oncology trials**. Stat Med, *34*(13), 2128–2137.

Eubank, R. L. (1988). **Spline smoothing and nonparametric regression**, volume 90. M. Dekker, New York.

Fisher, R. (1925). **Statistical methods for research workers**. Oliver and Boyd, Edinburgh.

Giaquinta, M., Hildebrandt, S. (1996). **Calculus of Variations 1. The Lagrangian Formalism**. Springer, Berlin.

Gleixner, A., Bastubbe, M., Eifler, L., Gally, T., Gamrath, G., Gottwald, R. L., Hendel, G., Hojny, C., Koch, T., Lübbecke, M. E., Maher, S. J., Miltenberger, M., Müller, B., Pfetsch, M. E., Puchert, C., Rehfeldt, D., Schlösser, F., Schubert, C., Serrano, F., Shinano, Y., Viernickel, J. M., Walter, M., Wegscheider, F., Witt, J. T., Witzig, J. (2018). **The SCIP optimization suite 6.0**. ZIB-Report 18-26, Zuse Institute Berlin. [acessed 2020-01-07].
URL `http://nbn-resolving.de/urn:nbn:de:0297-zib-69361`

GNU Project (2020). **GLPK (GNU linear programming kit)**. [accessed 2020-03-02].
URL `https://www.gnu.org/software/glpk/`

Gurobi (2018). **Mixed-Integer programming (MIP) basics**. [accessed 2018-03-03].
URL `http://www.gurobi.com/resources/getting-started/mip-basics`

Hyman, D. M., Puzanov, I., Subbiah, V., Faris, J. E., Chau, I., Blay, J.-Y., Wolf, J., Raje, N. S., Diamond, E. L., Hollebecque, A., *et al.* (2015). **Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations**. N Engl J Med, *373*(8), 726–736.

Ivanova, A., Paul, B., Marchenko, O., Song, G., Patel, N., Moschos, S. J. (2016). **Nine-year change in statistical design, profile, and success rates of phase II oncology trials.** J Biopharm Stat, *26*(1), 141–149.

Jeffreys, H. (1946). **An invariant form for the prior probability in estimation problems**. Proc Math Phys Eng Sci, *186*(1007), 453–461.

Jeffreys, H. (1998). **The theory of probability**. Oxford University Press, Oxford.

Jennison, C., Turnbull, B. W. (2000). **Group sequential methods with applications to clinical trials**. Chapman & Hall, London.

Jennison, C., Turnbull, B. W. (2015). **Adaptive sample size modification in clinical trials: start small then ask for more?** Stat Med, *34*(29), 3793–3810.

Jung, S.-H., Kim, K. M. (2004). **On the estimation of the binomial probability in multistage clinical trials**. Stat Med, *23*(6), 881–896.

Kieser, M., Englert, S. (2015). **Performance of adaptive designs for single-armed Phase II oncology trials**. J Biopharm Stat, *25*(3), 602–615.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C. (2016). **Jupyter Notebooks – a publishing format for reproducible computational workflows**. In F. Loizides, B. Schmidt, editors, **Positioning and Power in Academic Publishing: Players, Agents and Agendas**, pp. 87 – 90. IOS Press.

Kröger, O., Coffrin, C., Hijazi, H., Nagarajan, H. (2018). **Juniper: an open-source nonlinear branch-and-bound solver in Julia**. In **Integration of constraint programming, artificial intelligence, and operations research**, pp. 377–386. Springer, New York.

Kunzmann, K. (2020a). **bad**. doi:10.5281/zenodo.3931867. Software package.

Kunzmann, K. (2020b). **badr**. doi:10.5281/zenodo.3625139. Software package.

Kunzmann, K. (2020c). **Examples for optimal adaptive designs for early phase II trials in clinical oncology**. doi:10.5281/zenodo.3625162.

Kunzmann, K., Kieser, M. (2016). **Optimal adaptive two-stage designs for single-arm trial with binary endpoint**. arXiv preprint arXiv:160500249.

Kunzmann, K., Kieser, M. (2017a). **Point estimation and p-values in phase II adaptive two-stage designs with a binary endpoint**. Stat Med, *36*(6), 971–984.

Kunzmann, K., Kieser, M. (2017b). **Test-compatible confidence intervals for adaptive two-stage single-arm designs with binary endpoint**. Biom J, *60*(1), 196–206.

Kunzmann, K., Kieser, M. (2020a). **Optimal adaptive single-arm phase II trials under quantified uncertainty**. J Biopharm Stat, *30*(1), 89–103.

Kunzmann, K., Pilz, M., Herrmann, C. (2020b). **adoptr**. doi:10.5281/zenodo.2616951. Software package.

Kunzmann, K., Pilz, M., Herrmann, C., Rauch, G., Kieser, M. (2020c). **The `adoptr` Package: Adaptive Optimal Designs for Clinical Trials in `R`**. J Stat Softw, *accepted*.

Lancaster, H. O. (1961). **Significance tests in discrete distributions**. J Am Stat Assoc, *56*(294), 223–234.

Lee, J., Leyffer, S. (2011). **Mixed integer nonlinear programming**, volume 154. Springer Science & Business Media, New York.

Lehmacher, W., Wassmer, G. (1999). **Adaptive sample size calculations in group sequential trials**. Biometrics, *55*(4), 1286–1290.

Liu, G. F., Zhu, G. R., Cui, L. (2008). **Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval**. Stat Med, *27*(4), 584–596.

Mander, A., Thompson, S. (2010). **Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials**. Contemp Clin Trials, *31*(6), 572–578.

Müller, H.-H., Schäfer, H. (2004). **A general statistical principle for changing a design any time during the course of a trial**. Stat Med, *23*(16), 2497–2508.

Nelder, J. A., Mead, R. (1965). **A simplex method for function minimization**. Comput J, *7*(4), 308–313.

Nemhauser, G., Wolsey, L. (1988). **Integer and combinatorial optimization**. John Wiley & Sons, Hoboken, New Jersey.

O'Hagan, A., Stevens, J. W., Campbell, M. J. (2005). **Assurance in clinical trial design**. Pharm Stat, *4*(3), 187–201.

Park, S. Y., Bera, A. K. (2009). **Maximum entropy autoregressive conditional heteroskedasticity model**. J Econom, *150*(2), 219–230.

Pilz, M., Kunzmann, K., Herrmann, C., Rauch, G., Kieser, M. (2019). **A variational approach to optimal two-stage designs**. Stat Med, *38*(21), 4159–4171.

Proschan, M. A., Hunsberger, S. A. (1995). **Designed extension of studies based on conditional power**. Biometrics, *51*(4), 1315–1324.

R Core Team (2019). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. [accessed 2019-12-03]. URL `https://www.R-project.org/`

Ragan-Kelley, B., Willing, C. (2018). **Binder 2.0 – reproducible, interactive, sharable environments for science at scale**.

Rufibach, K., Heinzmann, D., Monnet, A. (2020). **Integrating phase 2 into phase 3 based on an intermediate endpoint while accounting for a cure proportion—With an application to the design of a clinical trial in acute myeloid leukemia**. Pharm Stat, *19*, 44–58.

Salsburg, D. (2001). **The lady tasting tea**. Henry Holt and Company, New York.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., Neuenschwander, B. (2014). **Robust meta-analytic-predictive priors in clinical trials with historical control information**. Biometrics, *70*(4), 1023–1032.

Schoenfeld, D. (1981). **The asymptotic properties of nonparametric tests for comparing survival distributions**. Biometrika, *68*(1), 316–319.

Sertkaya, A., Birkenbach, A., Berlind, A., Eyraud, J. (2014). **Examination of clinical trial costs and barriers for drug development**. Technical report, Eastern Research Group, Inc., Lexington. [accessed 2020-01-31].
URL https://aspe.hhs.gov/report/examination-clinical-trial-costs-and-barriers-drug-development

Shan, G., Wilding, G. E., Hutson, A. D., Gerstenberger, S. (2016). **Optimal adaptive two-stage designs for early phase II clinical trials**. Stat Med, *35*(8), 1257–1266.

Simon, R. (1989). **Optimal two-stage designs for phase II clinical trials**. Control Clin Trials, *10*(1), 1–10.

Spiegelhalter, D. J., Freedman, L. S. (1986). **A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion**. Stat Med, *5*(1), 1–13.

Therasse, P., Arbuck, S. G., Eisenhauer, E. A., Wanders, J., Kaplan, R. S., Rubinstein, L., Verweij, J., Van Glabbeke, M., van Oosterom, A. T., Christian, M. C., Gwyther, S. G. (2000). **New guidelines to evaluate the response to treatment in solid tumors**. J Natl Cancer Inst, *92*(3), 205–16.

Thomas, D. W., Burns, J., Audette, J., Carroll, A., Dow-Hygelund, C., Hay, M. (2016). **Clinical development success rates 2006-2015**. Technical report, BIO. [accessed 2019-12-16].
URL https://www.bio.org/sites/default/files/legacy/bioorg/docs/Clinical%20Development%20Success%20Rates%202006-2015%20-%20BIO,%20Biomedtracker,%20Amplion%202016.pdf

U.S. Food and Drug Administration (1997). **ICH E8, general considerations for clinical trials**. Technical report, U.S. Food and Drug Administration. [accessed 2019-04-20].
URL https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-8-general-considerations-clinical-trials-step-5_en.pdf

U.S. Food and Drug Administration (1998). **ICH E9, statistical principles for clinical trials**. Technical report, U.S. Food and Drug Administration. [accessed 2019-04-20].
URL https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf

U.S. Food and Drug Administration (2018). **Fast track, breakthrough therapy, accelerated approval, and priority review - accelerated approval**. [accessed 2018-01-13].
URL https://www.fda.gov/ForPatients/Approvals/Fast/ucm405447.htm

Vandemeulebroecke, M. (2006). **An investigation of two-stage tests**. Stat Sin, *16*, 933–951.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., *et al.* (2014). **Use of historical control data for assessing treatment effects in clinical trials**. Pharmaceutical statistics, *13*(1), 41–54.

Wasserstein, R. L., Lazar, N. A., *et al.* (2016). **The ASA's statement on p-values: context, process, and purpose**. Am Stat, *70*(2), 129–133.

Wassmer, G., Brannath, W. (2016). **Group sequential and confirmatory adaptive designs in clinical trials**. Springer, Heidelberg.

Wolsey, L. A., Nemhauser, G. L. (1999). **Integer and combinatorial optimization**, volume 55. John Wiley & Sons, Hoboken, New Jersey.

18. Related Own Publications

Kunzmann, K., Pilz, M., Herrmann, C., Rauch, G., Kieser, M. (2020). **The `adoptr` Package: Adaptive Optimal Designs for Clinical Trials in `R`**. J Stat Softw, *accepted*.

This manuscript describes the `R` package `adoptr` which implements the ideas developed by Maximilian Pilz and myself in collaboration with Prof. Kieser, Prof. Rauch, and Caroline Herrman. The software was used in Section 13. I lead the development of the `R` package itself and provided guidance to Maximilian Pilz who contributed equally to its completion. I also lead the writing of the manuscript in close collaboration with the other authors.

Kunzmann, K., Kieser, M. (2020). **Optimal adaptive single-arm phase II trials under quantified uncertainty**. J Biopharm Stat, *30*(1):89-103.

This manuscript discusses some ideas related to optimisation under uncertainty and was partly used as basis for Chapter 3. I developed the core concept of optimisation under uncertainty, implemented the methods, and lead the writing of the manuscript in close collaboration with Prof. Kieser.

Pilz, M., Kunzmann, K., Herrmann, C., Rauch, G., Kieser, M. (2019). **A variational approach to optimal two-stage designs**. Stat Med, *38*(21):4159-4171.

This manuscripts discusses the principles of addressing the problem of optimal two-stage designs for continuous outcomes using variational methods. The ideas discussed in Chapter 7 are largely based on consideration leading to this manuscript. I proposed the formulation as variational problem, the solution approach via the Euler-Lagrange equation, supported the implementation, and the writing of the manuscript. Maximilian Pilz took the lead in making the approach work in practice and the writing of the manuscript in close collaboration with the other authors.

Kunzmann, K., Kieser, M. (2018). **Test-compatible confidence intervals for adaptive two-stage single-arm designs with binary endpoint**. Biom J, *60*(1):196-206.

The manuscripts builds on the idea of test-compatibility and explores some alternative interval estimators. Parts of it have been used as basis for the discussion of confidence intervals and credible intervals in Section 5.2. I developed the concept, implemented the methodology, and lead the writing of the manuscript in close collaboration with Prof. Kieser.

Kunzmann, K., Kieser, M. (2017). **Point estimation and p-values in phase II adaptive two-stage designs with a binary endpoint.** *Stat Med*, *36*(6):971-984.

This manuscript lays the basis for the discussion of frequentist inference methods in Section 5.2. I developed the concept of test-compatibility, implemented the methodology, and lead the writing of the manuscript in close collaboration with Prof. Kieser.

Kunzmann K., Kieser M. (2016). **Optimal adaptive two-stage designs for single-arm trial with binary endpoint**. arxiv:1605.00249.

This manuscript was the first to outline a simplified variant of the integer linear programming approach described in Section 2.3. I developed the core concept and the implementation, the manuscript was written jointly with Prof. Kieser who also contributed greatly to the choice of examples, the discussion and the introduction besides providing general guidance on the structure of the paper.

A. Appendix

A.1. Software

The overall solution speed of the problems crucially depends on two factors (see Section 8.1 and Figure 8.2). Firstly, the formulated integer linear problems (see Section 2.3) must be solved reliably and quickly. This can be done by any suitable (M)ILP solver. Commercial solvers like Gurobi (Gurobi, 2018) tend to be more reliable and faster but suffer from restrictive licensing. Therefore, the open-source GNU Linear Programming Kit (GLPK) (GNU Project, 2020) is used throughout this thesis.

The second determining factor for overall solution speed is the time required to formulate specific problem instances and pass them to the solver. Even for small problems the number of variables quickly exceeds several thousand and medium-sized problems might contain as many as 500 000 individual binary variables. Computing the coefficients of the various constraints typically requires excessive looping which is ineffective in scripting languages like `R` (R Core Team, 2019). `Julia` (Bezanson *et al.*, 2017) is a relatively new programming language that is ideally suited for this task. It achieves similar performance to statically compiled `C` while providing a high-level interface comparable to `R`. Within `Julia`, the `JuMP` package for mathematical optimisation (Dunning *et al.*, 2017) is a very convenient tool for implementing integer linear problems of the class discussed in this thesis. `JuMP` implements a solver-agnostic interface such that different solvers can easily be used with the same problem.

All methods described in this thesis are implemented in the publicly available `Julia` package 'bad' (Binary Adaptive Designs) (Kunzmann, 2020a).

While `Julia` is a powerful language it is not yet commonly adopted in the statistical community. `R` (R Core Team, 2019) is still much more wide-spread. To overcome this limitation, a thin wrapper `R`-package, `badr`, was implemented (Kunzmann, 2020b). This package enables users to interact with the `Julia` package `bad` through `R`. The only prerequisite is a working installation of `Julia` which is readily available for all major platforms.

The methodology to optimise designs for continuous endpoints was developed jointly with Maximilian Pilz and is independently available in the `R`-package `adoptr` (Kunzmann *et al.*, 2020b,c).

A.2. Reproducibility of results

Special care was taken to ensure the reproducibility of the results presented in this thesis. All examples and the code to generate the plots and figures are available at zenodo.org Kunzmann (2020c). Zenodo.org is a long-term storage system for research data operated by CERN and all records are permanently accessible via their digital object identifiers. The repository contains a set of interactive 'Jupyter notebooks' (Kluyver *et al.*, 2016) that were used to generate all plots.

The notebooks can be explored interactively and without installation of any software using a so called 'Binder-link'. Binder (Ragan-Kelley *et al.*, 2018) is a tool that turns software dependency specification into Docker containers (for more information on Docker containers see `https://www.docker.com/`) in a reproducible way and makes them available through web links in the browser via the free web service `https://mybinder.org`. The containerisation approach ensures highest levels of reproducibility since the software environment in which the scripts to produce all figures and plots is exactly the same at each execution. It also makes it possible to explore the examples interactively without having to install any software. At the time of publication of this thesis, this service was offered free of charge although the computing power of each instance is severely limited and computations can take substantially longer than on a more performant machine. In particular,

- Section 1.2, Section 1.3, and Section 1.4 can be explored at:
  `https://mybinder.org/v2/gh/kkmann/optimal-binary-two-stage-designs/0.2.2?urlpath=lab/tree/notebooks/introduction.ipynb`
- Section 8.1 and Section 8.2 can be explored at:
  `https://mybinder.org/v2/gh/kkmann/optimal-binary-two-stage-designs/0.2.2?urlpath=lab/tree/notebooks/optimal-two-stage-designs.ipynb`
- Section 9.1, Section 9.2, and Section 9.3 can be explored at:
  `https://mybinder.org/v2/gh/kkmann/optimal-binary-two-stage-designs/0.2.2?urlpath=lab/tree/notebooks/optimisation-under-uncertainty.ipynb`
- Section 10.1 can be explored at:
  `https://mybinder.org/v2/gh/kkmann/optimal-binary-two-stage-designs/0.2.2?urlpath=lab/tree/notebooks/bayesian-inference.ipynb`
- Section 11.1, Section 11.2, and Section 11.3 can be explored at:
  `https://mybinder.org/v2/gh/kkmann/optimal-binary-two-stage-designs/0.2.2?urlpath=lab/tree/notebooks/frequentist-inference.ipynb`
- Section 12.1 and Section 12.2 can be explored at:
  `https://mybinder.org/v2/gh/kkmann/optimal-binary-two-stage-designs/0.2.2?urlpath=lab/tree/notebooks/unplanned-adaptations.ipynb`

- Section 13 can be explored at:
  https://mybinder.org/v2/gh/kkmann/optimal-binary-two-stage-designs/0.2.2?urlpath=lab/tree/notebooks/continuous-case.ipynb

For more information on how to reproduce the results on a local machine, see the repository available at https://github.com/kkmann/optimal-binary-two-stage-designs or Kunzmann (2020c).

Finally, an interactive web application based on the 'shiny' framework https://shiny.rstudio.com/ is contained in the examples-repository Kunzmann (2020c). This application is also hosted using Binder and https://mybinder.org and can be accessed via https://mybinder.org/v2/gh/kkmann/optimal-binary-two-stage-designs/0.2.2?urlpath=shiny/shiny/.

A.3. Acknowledgements

I would like to thank Prof. Dr. sc. hum. Meinhard Kieser for his encouragement to write this dissertation at the Institute of Medical Biometry and Informatics in Heidelberg and the Deutsche Forschungsgemeinschaft for supporting my research with Grant KI-3. I am particularly grateful to Prof. Dr. rer. nat. Annette Kopp-Schneider for taking over the supervision of this thesis in a difficult situation.

I would also like to thank my colleges and former members of the institute for the friendly atmosphere and fruitful discussions. In particular, I want to express my gratitude towards Dr. Christian Stock, Dr. Katharina Hees, Dr. Laura Benner, and MSc. Maximilian Pilz.

I am deeply indebted to my sister, Dr. Nadine Pinder, whose thorough review of the manuscript helped me improve its final presentation in many ways and to my mother, Erika Kunzmann, for her encouragement and unfaltering support.

Last but not least, I want to thank Enzo and his team for supporting this dissertation with an inexhaustible supply of freshly distilled Italian coffee.

A.4. Affidavit - Eidesstattliche Versicherung

EIDESSTATTLICHE VERSICHERUNG

1. Bei der eingereichten Dissertation zu dem Thema 'Optimal Adaptive Designs for Early Phase II Trials in Clinical Oncology' handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum Unterschrift