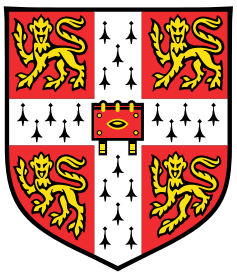
Curtailed phase II binary outcome trials and adaptive multi-outcome trials



Martin Law

Supervisors: Prof. Adrian P. Mander

Dr. Michael J. Grayling

MRC Biostatistics Unit

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Darwin College November 2020

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Martin Law

November 2020

Acknowledgements

I would like to acknowledge my supervisors, Prof Adrian P. Mander and Dr Michael J. Grayling, professionally for their excellent suggestions, direction and proof-reading, which made this thesis possible. Personally, I thank them for their constant commitment to a student hundreds of miles away, their positive attitude and above all their endless patience.

I would like to acknowledge the MRC Biostatistics Unit, for funding this work.

Finally, I would like to acknowledge Lisa, who looked after all the animals and me when I wasn't able to.

Abstract

Phase II clinical trials are a critical aspect of the drug development process. With drug development costs ever increasing, novel designs that can improve the efficiency of phase II trials are extremely valuable.

Phase II clinical trials for cancer treatments often measure a binary outcome. The final trial decision is generally to continue or cease development. When this decision is based solely on the result of a hypothesis test, the result may be known with certainty before the planned end of the trial. Unfortunately though, there is often no opportunity for early stopping when this occurs.

Some existing designs do permit early stopping in this case, accordingly reducing the required sample size and potentially speeding up drug development. However, more improvements can be achieved by stopping early when the final trial decision is very likely, rather than certain, known as stochastic curtailment. While some authors have proposed approaches of this form, these approaches have limitations, such as relying on simulation, considering relatively few possible designs and not permitting early stopping when a treatment is promising.

In this thesis we address these limitations by proposing design approaches for single-arm and two-arm phase II binary outcome trials. We use exact distributions, avoiding simulation, consider a wider range of possible designs and permit early stopping for promising treatments. As a result, we are able to obtain trial designs that have considerably reduced sample sizes on average.

Following this, we switch attention to consider the fact that clinical trials often measure multiple outcomes of interest. Existing multi-outcome designs focus almost entirely on evaluating whether all outcomes show evidence of efficacy or whether at least one outcome shows evidence of efficacy. While a small number of authors have provided multi-outcome designs that evaluate when a general number of outcomes show promise, these designs have been single-stage in nature only. We therefore propose two designs, of group-sequential and drop the loser form, that provide this design characteristic in a multi-stage setting. Previous such multi-outcome multi-stage designs have allowed only for a maximum of two outcomes; our designs thus also extend previous related proposals by permitting any number of outcomes.

Table of contents

Nomenclature

CP Conditional power

DtL Drop the loser

ENM Expected number of measurements

ESS Expected sample size

IQR Inter-quartile range

LFC Least favourable configuration

MAMS Multi-outcome multi-stage

MUE Median unbiased estimator

MVN Multivariate normal distribution

NSC Non-stochastic curtailment

RMSE Root mean squared error

SC Stochastic curtailment

SPRT Sequential probability ratio test

UMVUE Uniformly minimum variance unbiased estimator

Trial design glossary (selected)

Design approach/type: A way of designing a trial, for example, Simon's design.

Design realisation: A fully characterised example of a design approach/type, for example Simon's design with a particular set of values describing the number of participants at each analysis and the required number of responses at each analysis.

Feasible design (realisation): A design realisation that satisfies some required type-I error-rate and power.

Optimal design (realisation): A feasible design realisation that is superior to all other feasible design realisations when there is only a single criterion of interest.

Admissible design (realisation) A feasible design realisation that is superior to all other considered feasible designs for some weighted combination of multiple optimality criteria.

Dominated design (realisation): A feasible design realisation that is not superior to all other considered feasible design realisations with regards to any weighted combination of multiple optimality criteria.

Omni-admissible design (realisation): An admissible design realisation that is superior to all other admissible design realisations across all design approaches examined, for some weighted combination of multiple optimality criteria. An omni-admissible design realisation therefore has the smallest expected loss of all admissible design realisations across all design approaches examined, for a loss function with a given weighting.

(p_0, p_0) Response rates on control and treatment arms for which the type-I error-rate is controlled (two-arm)

(p_0, p_1) Response rates on control and treatment arms for which the type-II error-rate is controlled (two-arm)

α Type-I error-rate

β Type-II error-rate

$\boldsymbol{\delta}_0$ Vector of anticipated lower effect sizes for each outcome

$\boldsymbol{\delta}_1$ Vector of anticipated upper effect sizes for each outcome

$\boldsymbol{\delta}_\beta$ Vector of anticipated effect sizes for which a trial is powered

Δ Parameter used in calculation of stopping boundaries

$\hat{\tau}_{jk}$ Observed effect for outcome k at stage j

$\mathbb{I}$ Indicator function

$\mathbf{e}$ Vector of trial upper stopping boundaries

$\mathbf{f}$ Vector of trial lower stopping boundaries

$\mathscr{I}_j$ Information level at stage j

$\mathscr{R}$ Set of all $\{r, N\}$ or $\{r_1, n_1, r, N\}$

$\mathscr{T}$ Set of all terminal points

μ_k Effect size of outcome k

$\rho_{k_1 k_2}$ Correlation coefficient of outcomes k_1, k_2

σ_k^2 Variance of outcome k

Θ Maximum permitted number of CP values

θ CP values for a particular set $\{r,N\}$ or $\{r_1,n_1,r,N\}$

θ_E Upper threshold for CP values, above which a trial will stop for a go decision

θ_F Lower threshold for CP values, below which a trial will stop for a no go decision

$\theta_{E_{MIN}}$ Lower limit for θ_E in a design search

$\theta_{F_{MAX}}$ Upper limit for θ_F in a design search

B Number of participants per randomised block, per arm

C Constant used in calculation of stopping boundaries

CP_L, CP_U Interim lower and upper bounds for outcomes, in terms of conditional power (DtL design)

H_0 Null hypothesis

H_1 Alternative hypothesis

J Maximum number of stages

K Total number of outcomes

K_{max} Maximum number of outcomes that may be retained at the interim analysis (Dtl design)

m (multiple outcome design) Number of outcomes for which a trial is powered to show promise

m (single outcome design) Number of participants so far

m_C Number of participants so far on control arm

m_T Number of participants so far on treatment arm

N Total sample size

n Sample size per stage (multiple outcome)

N_j Total sample size at stage j

n_j Sample size at stage j

p Response rate (single-arm)

p_0 Response rate for which the type-I error-rate is controlled (single-arm)

p_1 Response rate for which the type-II error-rate is controlled (single-arm)

p_C Response rate on control arm (two-arm)

p_T Response rate on treatment arm (two-arm)

r Final rejection boundary

$S(m)$ Number of responses (single-arm) or number of successes (two-arm) after m participants

w_0, w_1 Weights of multiple optimality criteria

X_C Number of responses on control arm

X_T Number of responses on treatment arm

Z_{jk} Test statistic for outcome k at stage j

Chapter 1

Introduction

A clinical trial is an evaluation of one or more treatments for a medical condition. Such treatments may take a wide range of forms, from surgery to using a mobile phone application. The main purpose of a clinical trial may depend on what is already known about a treatment, and is often determined by the "phase" of the trial. In human trials there are four phases, and typically the main purpose of each trial phase is as follows: phase I focuses on safety, toxicity and finding the optimum dose; phase II focuses on determining if there is evidence that the treatment has the intended effect on the condition; phase III is similar to phase II, but on a larger scale and in comparison to an existing treatment; phase IV trials are larger still, and seek to identify side effects normally too rare to be found in smaller trials. However, a new treatment may be developed without strict adherence to the above concept of four separate phases (and associated trials). Whatever the purpose of a particular clinical trial, clinical trials in general are the means through which new treatments are evaluated.

1.1 Single-arm binary outcome designs

1.1.1 Motivation

Most novel treatments are found to be inefficacious, which makes the average development cost associated with each successful treatment extremely high [1]. Furthermore, trials themselves are expensive to run [2], and the nature of evaluating treatment response in oncology trials means that results are not immediately available, meaning that trials can take substantial time to complete. This makes novel designs that can improve the efficiency of clinical research extremely valuable.

Phase II clinical trials for cancer treatments often have a binary primary outcome, based on change in tumour size as measured by the RECIST criteria [3], and typically contain only a single arm. The aim of such a phase II trial is to gain enough information to decide whether a treatment should be carried forward for further testing (a go decision) or abandoned (a no go decision). In general, if a sufficient number of (positive) responses are observed, a go decision is made and some corresponding null hypothesis is rejected, otherwise a no go decision is made and the corresponding null hypothesis is not rejected. The most simple design to evaluate a treatment with a binary outcome is the single-stage design, described by A'Hern [4]. In a single-stage design, a fixed number of participants are recruited and once the trial is completed, a go or no go decision is made based on the number of responses. A single-stage design can be characterised by just two numbers: the number of participants to recruit and the number of responses required to make a go decision. When using a single-stage design, there is no opportunity to reduce the sample size, even if the final (go or no go) decision is known with certainty long before the planned end of the trial.

A number of designs have been proposed that can reduce the expected sample size (ESS) of a single-arm binary outcome trial compared to a single-stage design. Some the designs most relevant to this thesis are described below.

1.1.2 Existing designs

1.1.2.1 Simon

Simon's design [5] is the most frequently used phase II design amongst UK Clinical Trials Units, and with the exception of the standard single-stage trial, is the most frequently used phase II oncology trial design across the world [6, 7]. Simon's design is a two-stage design, meaning that in addition to the final analysis, it also contains one interim analysis. At a trial interim analysis, the current trial results are noted and some decision is made with respect to the trial. In this thesis, the only decision we consider is whether a trial should stop (to make either a go or no go decision) or continue. In Simon's design, the interim analysis takes place once a specified number of results are available. At this point, the trial stops for a no go decision if the number of responses is not greater than some specified value, otherwise the trial continues, recruiting the remaining participants and continuing until results are available for all participants. A go decision is permitted only if the trial continues to this second stage, and after all results are available. Compared to the single-stage design, two further values are required to describe Simon's design: the number of participants at which the interim analysis takes place, and the number of responses required at the interim analysis to continue the trial. The possibility of making a no go decision at this interim point, before the end of the trial, has the effect of reducing the ESS compared to a single-stage design [5].

1.1.2.2 Mander and Thompson

The design of Mander and Thompson [8], like Simon's design, contains a single interim analysis at which point the trial may end. However, at the interim analysis, Mander and Thompson not only allow stopping to make a no go decision, but also stopping to make a go decision if a specified number of responses (or more) has been observed. The number of responses required to make a go decision at the interim is greater than the number required to avoid a no go decision. This design retains the positive aspects of Simon's design [5] while

decreasing the ESS in the case that the treatment is efficacious. This design is characterised using one additional value compared to Simon's design: the number of responses required at the the interim analysis to make a go decision.

1.1.2.3 Chi and Chen

The design proposed by Chi and Chen [9] also consists of two stages. Like Simon's design, this design can also be described using four values: the final number of participants and required responses and an interim number of participants and required responses. The design differs from Simon's and Mander and Thompson's designs in the following aspects: the trial will end early to make a no go decision during the first stage as soon as it not possible to reach the number of responses required to continue at the interim analysis. That is, the design permits stopping in advance of the interim analysis. During the second stage, the trial will also end early to make a no go decision as soon as it is not possible to reach the final number of responses required (to make a go decision) at the end of the trial. That is, the design permits stopping between the interim and final analyses. Conversely, as soon as the observed number of responses reaches the number required at the interim analysis, a decision to proceed to the second stage will be made. Furthermore, the trial will end for a go decision as soon as the final number of required responses is reached. A distinction between this design compared to that of Mander and Thompson is that this design does not have a separate interim number of responses for stopping to make a go decision. Stopping a trial as soon as a specified number of responses can or can not be reached, used in this design, is known as *non-stochastic curtailment* (NSC). Consequently, we will refer to this design as the "NSC" design.

1.1.2.4 Stochastic curtailment and conditional power

It is possible to end a trial early not only when a go decision is either certain or no longer possible, as in NSC above, but also when a go decision is either likely or unlikely. This is known as *stochastic curtailment* (SC). A number of approaches are available for SC, three of which are described by Jennison and Turnbull [10]. One such approach is based on the concept of *conditional power* (CP). Conditional power (or "assumed conditional power" [11]) is the probability of rejecting some null hypothesis (and making a go decision), conditional on an anticipated treatment effect and the current number of participants and responses. The idea of CP can be used in conjunction with SC in the following way: if the CP is below some specified lower threshold, or exceeds some specified upper threshold, then a trial will end for a no go or go decision respectively. In this way, NSC can be seen as a special case of SC, where the lower threshold is equal to zero for stopping for a no go decision and, if permitted, the upper threshold is equal to one for stopping for a go decision. Two designs that incorporate SC using CP are described below.

1.1.2.5 Ayanlowo and Redden

Ayanlowo and Redden [12] propose an approach that is a direct extension of the single-stage design and Simon's design. Indeed, they describe their work as "examin[ing] the benefit of incorporating SC in... Simon's minimax and optimal designs". In common with Simon's design, the design does not permit stopping before the interim analysis and permits stopping only to make a no go decision. The design uses SC in the second stage, allowing a no go decision to be made if the CP of the trial decreases below a specified threshold. Ayanlowo and Redden examine two choices for this threshold, 0.05 and 0.10. A limitation of this approach is that the designs are found by obtaining either a single-stage or Simon design with a suitable type-I error-rate and power, then allowing early stopping due to SC. No other possible values for maximum sample size nor the interim or final required number of

responses is explored. This also means that these values are not altered to account for the possibility of early stopping. This can result in a decrease in both the type-I error-rate and power [12]. Ayanlowo and Redden obtain the ESS using simulations of size 1000, under the null hypothesis only. The CP is calculated solely on the probability of reaching the required number of responses at the end of the trial, and does not account for the increased probability of the trial ending early due to SC.

1.1.2.6 Kunz and Kieser

Kunz and Kieser [13] present a similar proposal to Ayanlowo and Redden [12]; that of incorporating SC to an existing design, either Simon's design or the NSC design. Again, SC is simply "added" to an existing design; the maximum sample size and interim and final required number of responses are not altered to account for the consequent increased probability of early stopping. Kunz and Kieser obtain the ESS using simulations of size 10,000. The approach of Kunz and Kieser is more general than Ayanlowo and Redden, for two reasons: firstly, SC is permitted at any point in the trial, rather than in the second stage only. Secondly, a uniform range $\{0, 0.01, \ldots, 1\}$ of lower thresholds for CP is examined, rather than just 0.05 and 0.10.

1.1.2.7 Further designs of interest

There are two additional design approaches worth categorising further, the characteristics of which will be used mostly in order to explain some aspect of the design search, the process by which designs are found. The first is the aforementioned single-stage work of A'Hern [4]. In particular, A'Hern provides both a range within which the final number of required responses of such a trial must exist, and also an equation for approximating the value itself. Further details are given in the relevant Methods section in Chapter 2.

The second design used to explain some aspect of the design search is the sequential probability ratio test (SPRT) of Wald [14]. This design contains upper and lower stopping boundaries for every possible number of participants, with accompanying equations for prescribing these provided. These upper and lower boundaries do not converge as the number of participants increases, and consequently the design has no maximum sample size. Again, further details are given in Chapter 2.

1.1.3 Continuous monitoring and informal curtailment

NSC and SC are typically described in terms of continuous monitoring, where the data are analysed after each participant's results become available. This may be considered a special case of sequential monitoring, which describes any trial in which interim results are analysed. Sequential analysis is methodologically well established [10, 15]. Continuous monitoring has been proposed not only in the single-arm approaches of Chi and Chen, Ayanlowo and Redden (second stage) and Kunz and Kieser, but also in designs for randomised binary outcome trials [16, 17]. In terms of practicality, continuous monitoring may be easier when a trial's recruitment rate is low [18], which is often the case in application: in a review of 122 trials, Campbell et al. [19] found that early participant recruitment was slower than expected in 77 (63%) of reviewed trials, and a review of 151 randomised controlled trials by Walters et al. [20] reported a median recruitment rate of 0.92 participants per centre per month. Furthermore, Campbell et al. [19] found that only 38 (31%) of 122 trials reached their intended sample size and 66 (54%) requested a trial extension. As such, trial recruitment rates are generally lower than anticipated, and may in some instances be low enough to facilitate continuous monitoring, especially when all stopping boundaries are obtained in advance and no additional statistical analysis is required to make a decision. Including interim analyses, and in particular continuous monitoring, in a clinical trial comes at an administrative and logistical cost [18], and for large trials, the potential savings in sample size may not outweigh

this cost. However, for small trials, continuous monitoring is attainable, as shown in the examples immediately below.

1.1.3.1 Continuous monitoring and informal curtailment: examples

Continuous monitoring may be expected to be specified at the trial design stage; see for example, Todd et al. [21] and McCabe et al. [22]. However, continuous monitoring and subsequent curtailment may also take place in trials where no such monitoring is specified in advance. In particular, authors may acknowledge the use of curtailment (and thus continuous monitoring) in a trial without using such terms in the corresponding manuscript. For example, Santana et al. [23] planned to follow a Simon design, but as soon as trial success was not possible, a no go decision was made. This resulted in a sample size saving of 33% ($n = 5$) compared to the planned Simon design. Necchi et al. [24] made a similar sample size saving (33%, $n = 7$) by using NSC before their interim analysis. Mego et al. [25] ended an optimal Simon design to make a no go decision, resulting in a sample size saving of 17% ($n = 3$), stating that "the study was terminated prematurely, because even if there were to be an objective response in the last 3 patients, the primary end point could not be reached". Furthermore, in the first stage of a Simon design, Wagner et al. [26] chose not to replace a patient who became inevaluable because it was not possible to reach the number of responses required to proceed to the second stage.

Informal curtailment is not restricted to stopping due to a lack of response. Stein et al. [27] conducted single-arm trials to test a treatment in two strata, using two Simon designs. Both trials were ended early because it was certain in both cases that the trials would end in success: "Informal analysis of these data (readily available to the lead investigator) indicated sufficient activity including complete responses to encourage further exploration of this regimen in either stratum." The total sample size saving was 27% ($n = 14$). Yu et al. [28] ended a planned Simon design early to make a go decision after 25 participants out of a

planned total of 55, less than halfway through. Moskwitz et al. [29] and Yoon et al. [30] also ended planned Simon designs early to make go decisions.

Informal curtailment may take place even when the response time is of considerable length. Using a primary endpoint of progression-free survival at 6 months, Sepúlveda-Sánchez et al. [31] stopped a Simon design early to make a no go decision, stating, "the study was closed at this point when the goal... could not be reached in the second stage". The sample size saving was 6% ($n = 2$). Similarly, Pedersen et al. [32] used a primary endpoint of overall survival at 6 months and ended their trial early "as the endpoint could not be reached", with a sample size saving of 20% ($n = 5$).

SC has also been used without being specified in advance. Odia et al. [33] conducted two concurrent trials, one of which required 4 responses out of 19. This trial was ended after observing 1 response out of 12 participants, "due to poor enrollment and therapeutic futility", clarifying in the discussion that "it is unlikely to find 3 objective responses among the remaining 7 patients". Thus the authors used SC, however informally. The sample size saving compared to the planned design was 37% $(n = 7)$.

Together, these examples show that curtailment through continuous monitoring is viable in practice, rather than purely theoretical, and thus methods that use continuous monitoring are valuable.

1.1.4 Inference

The above examples also suggest that continuous monitoring and subsequent curtailment is more common than citations of the associated methodological literature indicates. Whilst this is an important observation for motivating further design work in this area, an important additional consequence of unplanned monitoring that should be noted is that any resulting inference, such as that undertaken in the above examples, may result in biased point estimates and confidence intervals with low coverage [34, 35]. Accordingly, by not anticipating and

accounting for continuous monitoring and curtailment at the design stage, investigators are taking inferential risks. This lack of planning also has costs in terms of the ability to report accurate ESSs and trial duration at the design stage, which may directly affect expected financial costs and/or the ability to act quickly upon making a go or no go decision. These issues are amplified in the case of SC, where the ESS can be greatly reduced.

1.1.5 Single and multiple optimality criteria

A design *realisation* is a particular instance of a design, for example, a single-stage design where a go decision will be made if more than five responses are observed from a total of 20 participants. Changing either of these values would represent a different realisation. When choosing between a selection of design realisations, a method is required for choosing which is the "best". Typically this is done using an optimality criterion. Designs can be compared using a number of different optimality criteria, including their maximum sample size, ESS under certain anticipated response rates and ESS under certain anticipated response rates within the subset of designs that minimise the maximum sample size. A design realisation must also be *feasible*, that is, it must satisfy some chosen type-I error-rate and power. A feasible design that is the best-performing design realisation with respect to some single optimality criterion is known as the *optimal* design realisation for that criterion. The optimal design may differ depending on the criterion used and the design approach(es) considered.

In the setting of multiple optimality criteria, Jung et al. [36] determined the "best" design realisations by creating a loss function that was a weighted combination of two optimality criteria: maximum sample size and ESS under some null hypothesis. The authors describe a design realisation as *admissible* if it has the smallest expected loss of all considered design realisations, that is, it is superior to all other considered design realisations, for some weighted combination of optimality criteria. A particular design may be an admissible design for a range of combinations of weights. Our interest generally lies in finding the collection of

design realisations that comprise the admissible designs across all combinations of weights. A feasible design realisation that is not admissible, that is, not superior to any other feasible design realisations for any weighted combination of optimality criteria, is denoted *dominated*.

Mander et al. [37] extend the concept of the loss function and admissible designs by incorporating a third component to the loss function: ESS under an alternative hypothesis. Using a maximum of three weights means that all possible combinations of weights can be expressed in a triangle-shaped region, where the (x, y) co-ordinates represent two of the weights. The third weight is the complement of the sum of the first two weights. The triangle can then be divided into a grid of points. Each point on the triangle is given a colour, with the colour representing the design realisation with the lowest loss score, that is, the admissible design, for that particular combination of weights. This results in a set of admissible designs, each with its own region, covering all combinations of weights.

1.2 Randomised binary outcome designs

1.2.1 Motivation

Single-arm binary outcome phase II trials require fewer participants than equivalent randomised two-arm trials, making single-arm trials a pragmatic choice in many instances. The data from single-arm binary outcome trials are typically compared to a pre-specified historical control response rate. However, this comparison may not be valid [38–40]. For example, a systematic review of phase II oncology trials found that 46% ($N = 70$) of reviewed trials that used a historical response rate did not cite a source for the historical response rate used [38]. Two-arm randomised trials directly compare the responses of two groups from the same population, where one group has been allocated to treatment and the other group has been allocated to control, using randomisation. This is preferable to a non-randomised trial comparing the responses from a contemporary population to those from a historical

population, which may differ in characteristics. Although single-arm trials are more common in small populations and in rare disease settings, randomised designs should still be preferred when at all possible, as using historical information may provide less robust evidence [41, 42]. Thus, it has been argued that in almost all instances, two-arm randomised trials should be preferred to single-arm trials, with two-arm randomised trials considered to be the gold standard in trial design [43, 44]. Nevertheless, single-arm trials remain popular in phase II oncology, accounting for 57% of trials in a recent review of 557 trials [45]. It is therefore of interest to reduce sample sizes in two-arm phase II binary outcome trials, so that it may become possible to use two-arm designs when previously only a single-arm design would be considered, either due to cost or recruitment difficulties. For further details on the choice of single-arm or randomised designs, see Grayling et al. [39].

1.2.2 Existing two-arm designs

1.2.2.1 Jung

As discussed above, one approach to reducing sample size is to allow early stopping in the form of interim analyses. In the area of single-arm trials, the most frequently implemented such design is by Simon [5], as described in Section 1.1.2.1, where at a single interim analysis, the trial may stop early due to a lack of response. An analogous two-arm design has been proposed by Jung [46]; as with Simon's design, there is a single interim analysis and the trial may end at this point due to a lack of response. In this case, a lack of response entails observing a low response rate on the experimental treatment arm compared to the control arm.

1.2.2.2 Carsten and Chen

Further sample size savings can be made by allowing a trial to stop as soon as the final trial decision is certain, that is, by including NSC, with respect to either a positive effect or lack

of positive effect on the treatment arm compared to the control arm. This can be incorporated as an extension of Jung's design [46], by additionally allowing the trial to stop immediately if the response rates are such that the final go or no go decision is known with certainty, either at the interim (no go decision only) or by the end (go or no go decision). Equally, such a design may be understood as a two-arm extension of Chi and Chen's single-arm design [9]. This approach was first proposed by Carsten and Chen [17]. Their proposed design allows stopping after every pair of participants, where each pair is allocated evenly to the experimental treatment and control, therefore assuming perfect balance. As in Jung's design [46], an interim stopping boundary is used, which allows stopping only for a lack of response. "Success" is determined for every (balanced) pair of results, where success is defined as a pair of results where response on treatment and non-response on control is observed; all other combinations of results are treated equally, as a non-success. The test statistic is then the total number of such successes. The trial stops as soon as a pre-specified required number of successes is observed, or as soon as the number of successes required (either at the interim or at the end of the trial) cannot be reached.

1.2.2.3 Chen et al.

The incorporation of NSC into Jung's design is also proposed by Chen et al. [16], where the proposed design allows stopping after every patient. In this design, success is determined for each patient, and is defined as a response for a participant on treatment or a non-response for a participant on control. The test statistic in this case is the difference in the number of responses between the treatment and control arms; the same test statistic as is used by Jung [46]. The trial stops as soon as the required difference in the number of successes is guaranteed to be reached, or cannot be reached (again, either at the interim or by the end of the trial). Thus the designs of Carsten and Chen and Chen et al. require continuous monitoring [16, 17].

1.3 Multi-outcome multi-stage designs

In Chapter 4 we will consider single-arm trials where there are multiple key outcomes to account for, rather than one as in Chapter 2 and the designs outlined in Section 1.1. This is important as a clinical trial will typically measure many outcomes of interest. This may be done for a number of reasons. For example, investigators may plan to conduct a phase III trial using the outcome that will show the strongest evidence of treatment efficacy, and a multiple outcome trial in phase II will help identify this candidate outcome. Alternatively, investigators may wish to measure multiple outcomes in a phase III trial with the intention of declaring trial success where a promising treatment effect is observed on one of the outcomes. Furthermore, some disease conditions are typically assessed in a multi-dimensional manner, for example respiratory health [47]. There may also exist a core outcome set for the condition of interest, detailing a set of outcomes that should be measured when evaluating a treatment for that condition [48]. In general, there may simply be interest in observing a novel treatment's effects on a range of relevant endpoints.

1.3.1 Existing designs

1.3.1.1 Multi-outcome single-stage trials

A trial with multiple outcomes may be designed to evaluate whether a positive effect is observed on at least one of several key outcomes. Outcomes of this type are known as "multiple primary" outcomes [49]. The presence of multiple primary outcomes means that additional testing must be accounted for compared to the standard single-arm trials discussed in Chapter 2. Specifically, in the case of multiple primary outcomes, we must consider the family-wise error-rate, the probability of at least one type-I error occurring. For example, to control this as desired, we might apply a multiple testing correction such as the Bonferroni procedure.

In contrast to multiple primary outcomes, a multiple outcome trial may be designed to evaluate whether a positive effect is observed for all outcomes in some specified set. In this case, the outcomes are described as “co-primary" [49]. Sozu et al. have examined multiple outcome trials in detail, providing design approaches for both multiple and co-primary outcome trials [49].

A multi-outcome trial may be treated as having multiple null hypotheses, each of which may or may not correspond to a single outcome. Such trials may organise these hypotheses in a hierarchical manner. Furthermore, the type-I error-rate may be divided among these hypotheses and propagated from one to another or “spent” as hypotheses are rejected or not rejected. Maurer and Bretz describe such trials and extend this concept by introducing the idea of “memory”, whereby the order of propagation or spending of the type-I error-rate is taken into account [50]. Such approaches are beyond the scope of this work.

1.3.1.2 Group sequential trials

Group sequential trials, also known as multi-stage trials, include multiple interim analyses. Such trials may permit early stopping at these interim analyses if the current estimate of the treatment effect is greater than some upper boundary only, less than some lower boundary only, or either, where such boundaries are specified in advance [10]. Group sequential trials improve upon single-stage trials by permitting such early stopping at each interim analysis; typically this means that the expected (or average) sample size required by a multi-stage trial is below that required by the corresponding single-stage trial.

1.3.1.3 Multi-arm drop the loser trials

If more than one experimental treatment exists for a given condition of interest, then an improvement on undertaking a series of single-arm or two-arm trials is to use a multi-arm design. Multi-arm trials allow multiple experimental treatments to be simultaneously

compared to a common control treatment, reducing the required sample size compared to conducting a series of trials each with a single experimental treatment arm. The concepts of multi-arm and multi-stage trials can be combined to create a trial design containing multiple experimental treatment arms, tested over multiple interim analyses, or stages. Such trials are known as multi-arm multi-stage or MAMS trials.

In MAMS trials, where multiple experiment treatment arms are evaluated simultaneously, the number of arms may be reduced over the duration of the trial. This is typically undertaken by ceasing recruitment on an arm or arms, based on current results. This is known as "dropping" an arm (or arms). This provides a statistical advantage, as more participants can subsequently be recruited to the remaining, better-performing arms, providing more information about those arms. A disadvantage to dropping arms in general is that the required sample size is not fixed: a trial consisting of mostly well-performing treatments may have little or no dropping of arms. Conversely, a trial consisting of mostly poorly-performing treatments may drop many arms in the early stages. This makes the certain practical aspects of the design, such as trial duration and cost, uncertain.

One approach to dropping treatment arms is the "drop the loser" (DtL) design, where exactly one treatment arm is dropped at each stage or the number of arms at each stage is otherwise pre-determined [51]. As a result, the number of stages required for the best-performing treatment (or treatments) to reach some required number of participants is fixed and known in advance, even if the identity of that treatment (or those treatments) is not known in advance. This aspect of the DtL design allows MAMS trials to be planned with more certainty than other MAMS designs that do not have this property.

1.3.1.4 Dropping outcomes

Multi-outcome trials may continue to measure all planned outcomes even when some of the outcomes have a low probability of contributing to trial success, either in the multiple

primary or multiple co-primary outcome setting. If some outcomes are particularly expensive, invasive or time-consuming to measure, it may be valuable to use a trial design that stops measuring outcomes that are performing poorly. We describe this action as *dropping an outcome*, similar to the dropping of arms in the multi-arm setting.

1.3.1.5 Multi-outcome multi-stage trials

The approach of Sozu et al. to multiple primary and co-primary outcomes in clinical trials focuses on single-stage, two-arm designs, and they describe this work as a foundation for other design features, including group sequential trials [49]. A review by Hamasaki et al. of clinical trial designs that use co-primary endpoints describes numerous approaches to multi-outcome design, including multi-stage designs [52]. Among these are designs that include early stopping for a go decision only or for either a go or no go decision, for either two outcomes or two or more outcomes. These designs are classified in Table 1.1. In all cases, the designs use co-primary endpoints, and thus promising results must be observed on every endpoint for the null hypothesis to be rejected. There is no framework to test if some subset of outcomes show promising effects.

Author	Early stopping permitted	Number of outcomes
Ando et al.[53, 54]	Go decision only	2
Asakura et al.[55]	Go decision only	2
Cheng et al.[56]	Go decision only	2
Hamasaki et al.[57]	Go decision only	≥ 2
Cook and Farewell [58]	Go or no go decision	2
Jennison and Turnbull [59]	Go or no go decision	2
Schuler et al [60]	Go or no go decision	2
Asakura et al.[61]	Go or no go decision	≥ 2

Table 1.1 Summary of group sequential designs for co-primary outcomes by Hamasaki et al.

Other multi-outcome multi-stage designs exist beyond this summary, many of which relate to the specific sub-case of a single efficacy and single safety outcome. For example, in

the context of binary outcome trials, Conaway and Petroni present a single-arm, two-outcome, multi-stage design with co-primary outcomes, with the expectation that one outcome is an efficacy outcome and the other outcome is a safety outcome [62]. In this design, there are two type-I error-rates, one for each outcome, which may be set independently. Conaway and Petroni have also presented a similar two-stage design that incorporates a trade-off between efficacy and toxicity, allowing trial success for a lower than anticipated response rate if toxicity is low, and vice versa [63]. This approach is conceptually similar to the loss functions of Jung et al. and Mander et al. [36, 37]. Ristl et al. have proposed a multi-outcome design for two-arm binary outcome trials that allows rejection regions that have few constraints in terms of shape and that are optimal, where the definition of "optimal" may be specified [64]. In the area of multi-arm multi-stage trials, Jaki and Hampson propose a design whereby a single treatment arm is selected at the first interim analysis [65]. This treatment is selected based on a trade-off of one efficacy outcome and one safety outcome, both normally distributed, in addition to some minimum safety requirement. Thall and Cheng present a two-stage, two-outcome design, where trial success is based on a trade-off of the two outcome, again with the idea that one outcome pertains to efficacy and the other outcome to safety [66]. This design may be generalised to an arbitrary number of stages. Regarding efficacy and safety as binary outcomes, Bryant and Day propose a Simon-style two-outcome, two-stage design where a trial may end only for futility at the interim, when either the number of efficacy responses is low or the number of participants experiencing toxicity is high [67].

1.3.1.6 Separate vs. simultaneous stopping

Outside the case of multiple primary endpoints, where only a single outcome must show promise for the trial to be a success, a decision must be made regarding when to conclude that the necessary number of outcomes show promise. It is possible to conclude that an outcome is promising as soon as its test statistic is found to exceed a corresponding efficacy

stopping boundary. The outcome may cease to be measured at this point, which we describe as dropping an outcome, as described above (Section 1.3.1.4). Conversely, one may choose not to make conclusions regarding every outcome separately, but instead deem the trial a success only if enough outcome test statistics exceed their corresponding efficacy stopping boundaries simultaneously. In the multi-outcome case, this choice has been discussed previously [55, 57, 68]. This can be viewed as related to the two options for stopping MAMS trials, known as *separate* versus *simultaneous* stopping [69, 70]. In the area of MAMS, this choice is generalised by Grayling et al. [71], to permit stopping after a specified number of arm-specific null hypotheses have been rejected.

1.3.1.7 Number of outcomes required to show promise for trial success

The review of multi-outcome designs by Hamasaki et al. [52] covers co-primary endpoints only, where the trial is a success only if a certain degree of efficacy is shown on all measured outcomes. Sozu et al. [49] describe designs for co-primary endpoints and for multiple primary endpoints, where trial success is declared if a promising treatment effect is observed for any measured outcome. Historically, the focus of multi-outcome design is on co-primary endpoints and multiple primary endpoints. In contrast, Delorme et al. [72] describe a generalised power approach, a multi-outcome single-stage design where trial success is declared if some specified number of outcomes (or more), out of a larger set of outcomes, show promise. Mielke et al. have recently proposed two testing procedures for claiming trial success in this more general context, again in a single-stage trial, and allow outcomes of any type [73]. Mielke et al. use multiple hypotheses, one for each outcome. If some specified number of hypotheses (or more) are rejected, the trial is declared a success. The approaches we propose in Chapter 4 are also centred on this idea of declaring success if at least some specified number of outcomes, out of a larger set of outcomes, show promise.

1.3.1.8 Multi-arm multi-stage trials with generalised error-rates

In addition to Delorme et al.'s generalised power approach [72], Grayling et al. [71] present an approach to MAMS (rather than multi-outcome single-stage) trials that allows trials to be powered to find any specified number of efficacious arms. This design also features a generalised approach to stopping, permitting stopping as soon as any specified number of separate hypotheses are rejected.

1.3.1.9 Composite outcome trials

When multiple outcomes are to be measured, a choice must be made regarding whether or not to combine the measurements into a single composite outcome. Testing a single, composite measurement is statistically simpler, and may be appropriate when the outcomes are deemed suitable to combine [10]. However, both determining how to appropriately combine outcome measurements and interpreting the resulting composite outcome may still be difficult. Consequently, the multiple-outcome design may be preferred in the case that creating a composite outcome is either inappropriate or otherwise challenging. Composite designs may have multiple stages, and such a design will be examined in Chapter 4.

1.4 Thesis aims

The focus of this thesis is to present novel approaches to clinical trial design and in doing so, fill gaps in the literature. While the approaches are novel and have some practical or logistical burden with regards to implementation, these burdens do not exceed those of existing approaches that are described in this thesis. For example, SC and continuous monitoring has been proposed in single-arm binary outcome trials. In Chapter 2, we present an approach to SC that is novel: stopping for a go decision is permitted; the design search is more wide-ranging and exact distributions are used rather than simulation. While non-

stochastic curtailment has been previously proposed in randomised binary outcome trials, SC has not. In Chapter 3, we focus on a design approach that uses SC, and permits varied frequency of monitoring depending on the practical needs of the individual trial. In multi-stage trials with multiple outcomes, the existing literature focuses mostly on two-outcome trials and on declaring trial success only when promising effects are observed on at least one outcome or on all outcomes. Chapter 4 proposes two novel designs for multi-stage trials with multiple outcomes: one that permits any number of stages and one containing two stages and allows outcome measurement to cease at the interim analysis. Both designs allow any number of outcomes and permit allow specification of the number of promising effects that must be observed for trial success to be declared. The advantages and limitations of the designs are summarised, in Chapter 5, where recommendations for their use are also given.

1.5 Code

All original code to undertake this work has been written in *R* [74] and can be accessed online [75, 76].

Chapter 2

Novel stochastically curtailed designs for single-arm binary outcome phase II trials

This work is based on a paper by Law et al. [77], which has undergone peer review twice. In both reviews, the work was acknowledged as statistically sound but impractical. Consequently, we have increased our focus on the practical aspects of the work.

2.1 Methods

In this chapter, we present a novel generalised approach to SC in single-arm binary outcome trials. Two designs are proposed. One is a Simon-type design that allows SC after each participants' results, for either a go or no go decision, and contains an interim analysis in the design to which SC is added. The other design also allows SC after each participants' results, again for a go or no go decision, but has no interim analysis in the design to which SC is added. That is, it is an extension of the single-stage design described by A'Hern [4]. The

latter design will also be shown to be alterable to allow analyses that are less frequent than continuous monitoring, while still reducing the ESS.

Throughout the chapter, the experimental treatment is assumed to have a true response rate $p \in [0,1]$, that is, each patient outcome is assumed to be Bernoulli distributed with parameter p, $Bern(p)$. The sum of n independent and identically distributed Bernoulli random variables with parameter p follows a binomial distribution, $Binom(n,p)$. We test the null hypothesis $H_0 : p \leq p_0$ against the alternative hypothesis $H_1 : p > p_0$. For all designs, we use the notation $S(m)$ to denote the number of responses observed after m participants and $(S(m),m)$ to denote the point in a trial where $S(m)$ responses have been observed after m participants. We assume that results from participants are independent and identically distributed. Consequently, the number of responses observed at each stage are also independent. At any analysis in any trial described in this chapter, conducted using the first m participants data, the test statistic used to undertake the hypothesis test is the current number of responses $S(m)$. All single-arm binary outcome designs discussed here can be considered as simply a set of pairs of boundaries to be compared to this test statistic at certain points in the trial, with the boundaries themselves chosen to provide specified operating characteristics.

The trial is powered to a level $1-\beta$ under $p = p_1$, and the type-I error-rate is controlled to α when $p = p_0$. Available results [78] on the monotonicity of the power function in designs of the type considered here means that the type-I error-rate is then controlled to α over all of H_0 (i.e., for all $p \leq p_0$) and power is at least $1-\beta$ for all $p \geq p_1$. Commonly, the value of p_0 is chosen to be the greatest response rate that is deemed typical for standard of care, while p_1 is chosen to be the smallest response rate that is large enough to warrant further study.

2.1.1 Brief review of existing designs

For single-arm trials with a binary outcome, the most simple trial design is the single-stage trial described by A'Hern [4], which is comprised of N participants. This is deemed a success if the final number of responses $S(N)$ exceeds a specified boundary r.

Extending this approach, Simon's design [5] adds an interim analysis after n_1 participants, at which point the trial proceeds to recruit a further $N - n_1$ participants only if the number of responses is greater than a pre-specified r_1, otherwise it stops for a no go decision.

Mander and Thompson [8] proposed an extension to Simon's design, where the trial may additionally stop for a go decision after n_1 participants if the number of responses exceeds some upper boundary e_1.

For both single-stage and Simon's design, it is likely that the final go or no go decision is known before the termination of the trial: if the final number of required responses is reached after m participants, that is, $S(m) > r$, then the trial will be declared successful regardless of the data from the remaining participants. Conversely, if it is no longer possible to reach the final number of required responses after m participants, that is, $r + 1 - S(m) > N - m$, then the trial will be declared a failure, again regardless of the data from the remaining $N - m$ participants. Chi and Chen therefore extended Simon's design by permitting early stopping for a no go decision if it is certain that the required number of responses will not be reached; that is if $r_1 + 1 - S(m) > n_1 - m$, $m \leq n_1$ in the first stage or $r + 1 - S(m) > N - m$, $m > n_1$ in the second stage. Furthermore, and unlike Simon's design, the trial permits stopping early for a go decision. The trial stops early for a go decision as soon as the final required number of responses is reached, that is, $S(m) > r$. That is, the design uses NSC.

Ayanlowo and Redden and Kunz and Kieser [12, 13] took the concept of early stopping further, by allowing early stopping for a no go decision if the CP is below some lower threshold, that is, by allowing SC. Denote this lower threshold θ_F. Then, the specific values of θ_F examined by Ayanlowo and Redden and Kunz and Kieser were fixed (Ayanlowo and

Redden: $\theta_F \in \{0.05, 0.10\}$, Kunz and Kieser: $\theta_F \in \{0, 0.01, 0.02, \ldots, 1\}$). We discuss later exactly how the CP is calculated, in Section 2.1.7.

2.1.2 Limitations of existing designs

Some aspects of the above curtailed designs have scope for improvement. Firstly, the operating characteristics of curtailed designs have often been estimated using simulation. However, such estimates are subject to simulation error, with the exact distribution of each trial's possible outcomes remaining unknown.

Secondly, the approaches of Ayanlowo and Redden and Kunz and Kieser use fixed or uniformly distributed thresholds for CP, which may reduce the number of meaningful design realisations searched over.

Thirdly, in both Ayanlowo and Redden and Kunz and Kieser [12, 13], rather than taking curtailment into account when searching for the optimal design (for some definition of "optimal"), the optimal non-curtailed design is found and then SC is applied to it. This again means that a narrower range of possible designs is examined. Permitting SC only for a no go decision means that the probability of rejecting the null hypothesis decreases. The type-I error-rate and power are defined as the probability of rejecting the null hypothesis conditional on certain response rates, as described above. Thus a second consequence of the approach of Ayanlowo and Redden and Kunz and Kieser of applying SC for a no go decision only is that both the type-I error-rate and power may decrease relative to chosen design. This further reduces the number of possible design realisations, as many will not reach the required type-I error-rate and power once curtailment has been applied. Moreover, the design approaches of Ayanlowo and Redden and Kunz and Kieser give equations for evaluating CP, but these equations do not fully account for the early stopping caused by SC.

Finally, the designs detailed above do not have a generalised approach regarding the number of points at which early stopping is permitted: early stopping is permitted at a single

point, at any point during the the second stage or at any point for the duration of the trial. As a result, it may not be possible for an investigator to choose the degree of monitoring most appropriate for a given trial.

2.1.3 Proposed designs

We propose two designs where the trial may stop not only for a no go decision if the CP is below some lower threshold, that is, $CP < \theta_F$ but also for a go decision if the CP is greater than some upper threshold denoted θ_E, that is, $CP > \theta_E$. The first design we propose is a new type of stochastically curtailed two-stage design. This will be referred to as the "SC" design. The second design we propose can be understood as an otherwise single-stage design that allows stopping if a go decision becomes likely or unlikely. This will be referred to as the "m-stage" design.

Note that it will ultimately be possible for all go and no go decisions to be concatenated into N-length vectors of stopping boundaries $\mathbf{e} = (e_1, e_2, \ldots, e_N)$ and $\mathbf{f} = (f_1, f_2, \ldots, f_N)$ respectively. To fix the length of the vectors, regardless of the actual allowed timing of analyses, we may use $e_i = \infty$ and $f_i = -\infty$ at any points $i \in [1, N]$ where stopping is not permitted/possible. Thus, while we typically use different sets of terms to define each type of design, for example $\{r_1, e_1, n_1, r, N\}$ or $\{r, N, \theta_E, \theta_F\}$, and the values of these terms characterise a particular realisation of that design type, it is possible to characterise a realisation of any design type using $\mathbf{e}$ and $\mathbf{f}$ only. This is useful to recognise as it demonstrates why the comparisons that will be conducted between designs are fair: both previous and our newly proposed designs amount to methodologies for specifying $\mathbf{e}$ and $\mathbf{f}$. Viewed from this angle, our work focuses on enabling these boundaries to be chosen in a more flexible, efficient and logical manner than previous works.

The above fact regarding $\mathbf{e}$ and $\mathbf{f}$ is also useful as it means that for any single-arm design, all possible combinations of number of participants and responses so far can be represented

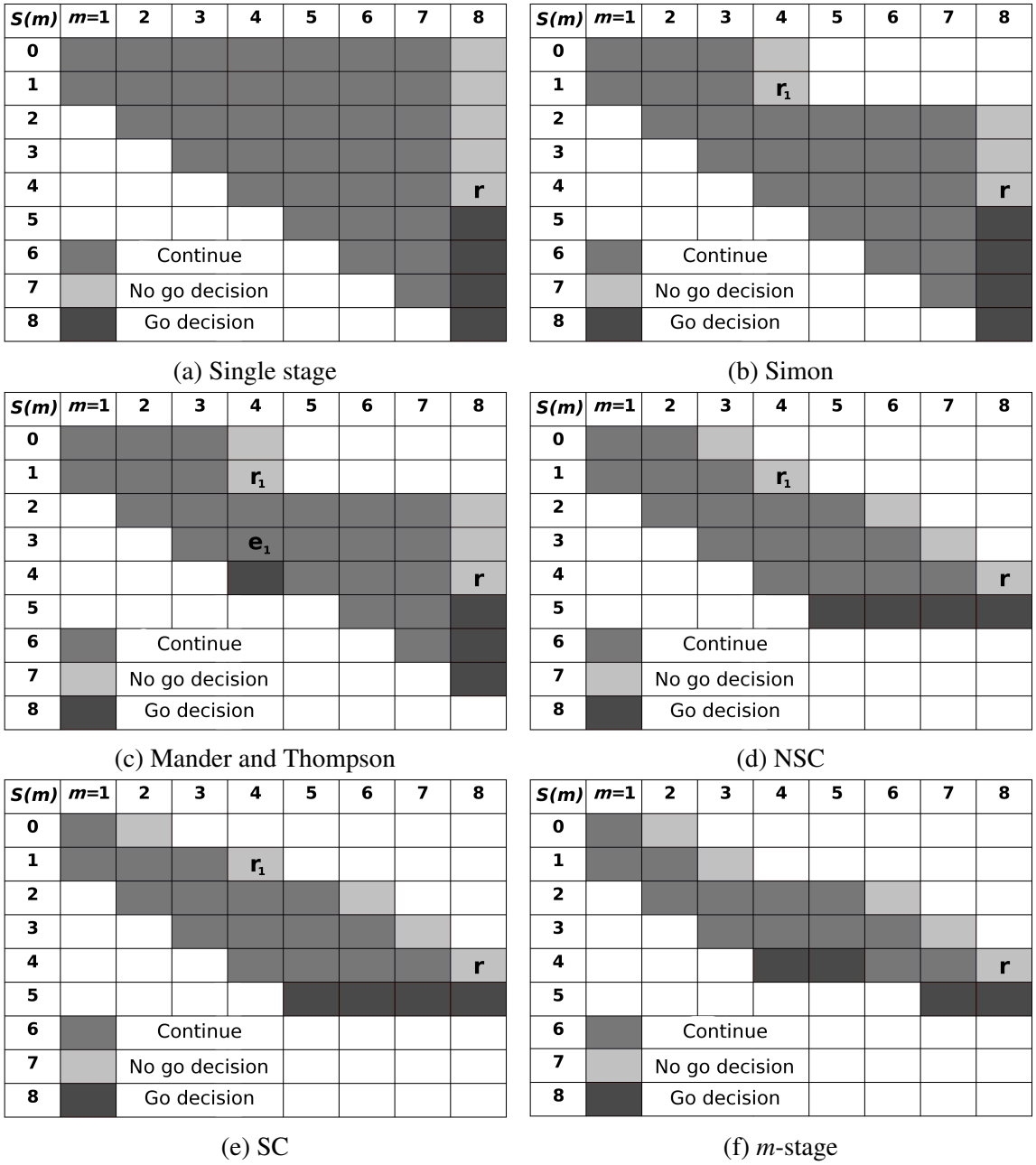
in an easy-to-understand grid. Examples of this are shown in Figure 2.1, for the following designs: single-stage, Simon, Mander and Thompson, NSC, SC and m-stage. Here, all possible points, that is, all possible participant and response combinations, and whether at each point the trial will continue or stop for a go or no go decision, are shown. This figure shows how these designs are related. It also shows in a practical sense how the incorporation of stochastic and non-stochastic curtailment can reduce the ESS by decreasing the number of points $(S(m),m)$ that can be reached.

2.1.4 Objections to curtailment

An objection to curtailment for a go decision could be that one would wish to obtain more data if a treatment appears promising. However, the current abundance of possible treatments to be tested among relatively few participants makes this argument less compelling than in the past.

There may be some objections to SC in particular, as it allows for the termination of a trial at a point where, under another design, the final decision is not yet certain. The primary rebuttal to this is to make it clear that the designs that will be proposed, in contrast to those of Ayanlowo and Redden and Kunz and Kieser, will retain the desired type-I error-rate and power; nothing is lost by using our approach to SC.

A more practical counterargument to objections to curtailment is that incorporating early stopping into Simon's design, using SC, is fundamentally no different incorporating early stopping in Simon's design compared to A'Hern's single-stage design. Furthermore, in Section 2.2, we use the example of a trial reported by Sharma et al. [79], who used Simon's design with $r_1 = 4, n_1 = 19, r = 15, N = 54$. CP is described more fully in Section 2.1.7 below, but briefly, the CP of a single-stage trial is $CP(p_1, S(m), m) = P(S(N) > r | p_1, S(m), m)$, with $p_1 = 0.4$ in this example. Consider a single-stage trial with the same sample size and final rejection boundary, $r = 15, N = 54$. At the points in the trial where Simon's design would



(a) Single stage

(b) Simon

(c) Mander and Thompson

(d) NSC

(e) SC

(f) m-stage

Fig. 2.1 Illustrative diagrams of different trial designs, showing potential paths where the study would end, known as terminal points. m: Number of participant results so far. $S(m)$: Number of responses so far. All trials have $N = 8, r = 4$, with $r_1 = 1$ in the two-stage designs and $e_1 = 3$ in Mander and Thompson's design. We may assume that (θ_F, θ_E) in the SC design are such that $(e_5, e_6, e_7, e_8) = (5,5,5,5)$, $(f_2, f_4, f_6, f_7, f_8) = (0,1,2,3,4)$ and that (θ_F, θ_E) in the m-stage design are such that $(e_4, e_5, e_7, e_8) = (4,4,5,5)$, $(f_2, f_3, f_6, f_7, f_8) = (0,1,2,3,4)$.

stop early, $S(19) \in \{0,1,2,3,4\}$, the CP of the single-stage trial is 0.30, 0.43, 0.56, 0.69 and 0.80 respectively. That is to say, if the trial reached the point $(S(m) = 4, m = 19)$, the conditional probability of rejecting H_0 would still be 0.80 for a treatment with response rate $p = p_1$ in the single-stage design, yet reaching this point requires stopping for a no go decision under Simon's design:

$$\begin{aligned} CP(p_1 = 0.4, S(m) = 4, m = 19) &= \mathrm{P}(S(54) > 15 | p = 0.4, S(19) = 4) \\ &= \mathrm{P}(S(N - n_1) > r - S(n_1) | p = 0.4) \\ &= \mathrm{P}(S(35) > 11 | p = 0.4) \\ &= \sum_{i=12}^{35} \binom{35}{i} 0.4^i (1 - 0.4)^{35-i} \\ &= 0.80, \text{ to 2 d.p.} \end{aligned}$$

Thus any acceptance of Simon's design is a tacit acceptance of ending a trial where, if a simpler design was employed, the final decision would not yet be certain.

2.1.5 Delayed responses

The trials we describe progress theoretically, and often practically, one participant result at a time, where each result is a response or non-response and each participant's result is known prior to any further enrollment. Such a trial is ideal in terms of minimising ESS. However, while a trial's recruitment rate may be low enough that participants are enrolled one at a time, it would be unusual for the results of all enrolled participants to be available prior to enrolling subsequent participants. The combination of recruitment rate and endpoint length, the length of time it takes to obtain a participant's result, has a direct effect on a trial's ESS. If indeed the endpoint length is short enough that the results of all currently enrolled

participants are always known before further enrollment, then the actual ESS will not differ from the calculated ESS. However, if, as expected in practice, the endpoint length is such that some results from currently enrolled participants are not known before further enrollment, then the actual ESS will be greater than the calculated ESS. The extent of this inflation of ESS will again depend on the combination of recruitment rate and endpoint length. If almost all results of enrolled participants are known at the point of further enrollment, the increase will be small. Conversely, if a low proportion of results of enrolled participants are known at the point of further enrollment, this increase may be considerable [80]. This is true of all designs using interim analyses of any kind, including curtailment. In our work, we assume that all participants' results are known before further recruitment, while acknowledging that inflation of ESS may occur in practice. Beyond a simple inflation in sample size, it may be the case that a decision is made based on currently available results but that such a decision is then contradicted by subsequent results. For example, current trial results may result in stopping early for a no go decision, but when combined with the subsequent results of currently enrolled participants, results indicate that the trial should continue. Such effects are beyond the scope of this work, though we acknowledge their gravity.

2.1.6 Obtaining exact distributions

There is a finite determinable number of possible sequences of participant results that lead to a point $(S(m), m)$; we consider each possibility as a "path" of a trial. As an example, see Figure 2.2: this example shows a possible path of a single-stage trial with no early stopping. This path may be described as $S(m) = 0,1,1,1,2,2,3,3$ for $m = 1,\ldots,N$. Note that a path may end at a point with $m < N$ if early stopping is permitted.

For a single-stage trial where no early stopping is allowed, the number of possible paths is 2^N. Recording the probabilities of all possible paths would allow the exact distribution of

S(m)	m=1	2	3	4	5	6	7	8
0								
1								
2								
3								
4								
5								
6								
7								
8		Trial path						

Fig. 2.2 Example of a path for a single-stage trial with $N = 8$.

the trial outcomes to be known. Even for a trial of moderate size, say $N = 30$, the number of paths $(2^{30} > 10^9)$ would make this computationally intensive.

However, to obtain the exact distribution of a trial, the probability of each path is not required: firstly, for any design, there are a number of points $(S(m), m)$ at which the trial would stop. Define these points "terminal points", and $\mathscr{T}$ as the set of all such points. The terminal points can be determined using the CP at each point in the trial. Secondly, it is not necessary to calculate the probability of reaching a given terminal point via each path, as many such paths have identical probabilities, reducing the computational burden of finding the probability of reaching a given terminal point.

The ESS for response rate p can then be obtained by multiplying the number of participants m at each terminal point by the probability of reaching that point:

$$ESS(p) = \sum_{m=1}^{N} \sum_{S(m)=0}^{m} \mathbb{I}[\{S(m), m\} \in \mathscr{T}] \, mU(S(m), m | p, \mathbf{e}, \mathbf{f}), \tag{2.1}$$

where $U(S(m),m|p)$ denotes the probability of reaching the point $(S(m),m)$ in a particular trial given true response rate p and vectors of stopping boundaries $\mathbf{e}$ and $\mathbf{f}$. Thus for any type of design, all that is required to find the ESS is the probability of reaching each terminal point in $\mathscr{T}$.

The sample size of a trial can also be described in terms of quantiles, including the median, in the following way: sort the sample sizes of the terminal points, that is, each m in each $\{S(m),m\} \in \mathscr{T}$, in ascending order. The corresponding probabilities $U(S(m),m|p,\mathbf{e},\mathbf{f})$ then comprise the cumulative density function, which can be used to calculate quantiles of the required sample size.

The type-I error-rate and power can also be obtained by summing the probabilities of reaching only the terminal points that result in a go decision under the assumptions $p = p_0$ and $p = p_1$ respectively. For example, power would be given by:

$$\sum_{m=1}^{N} \sum_{S(m)=0}^{m} \mathbb{I}[\{S(m),m\} \in \mathscr{T}]\mathbb{I}\{S(m) \geq e_m\} U(S(m),m|p,\mathbf{e},\mathbf{f}).$$

Moreover, it is possible to obtain the CP of a trial at any point $(S(m),m)$, as detailed below. Being able to obtain this information means that the exact distribution of the trial outcomes is known. In turn, the operating characteristics of the trial outcomes are known without recourse to simulation. This is valuable when searching for optimal designs, as simulation error could result in a sub-optimal design being chosen, whilst conducting a large number of simulations can also be time consuming.

2.1.7 Conditional Power

We define conditional probability $CP(p,S,m)$ as the probability of rejecting H_0 conditional on being at point $(S(m),m)$, when the true response rate is p. Setting $p = p_1$ gives the conditional power $CP(p_1,S,m)$. From here we refer only to conditional power rather than conditional probability and reiterate that "CP" is used to refer to conditional power. No

sample size re-estimation takes place. Strictly speaking, this is the "assumed conditional power". This is in contrast to the "observed conditional power", where the probability of rejecting the null hypothesis is conditioned on the maximum-likelihood estimate of the response rate using the current data. These two approaches are compared by Kunzmann et al. [11], who find assumed conditional power superior in terms of bias, mean absolute error and mean squared error when the true response rate is close to p_1, and conclude that the observed conditional power is "hard to justify theoretically". Ayanlowo and Redden and Kunz and Kieser [12, 13] also use the assumed conditional power, and previously provided equations for calculating CP, but these equations did not account for all early stopping due to SC.

For the NSC design, that is, stopping only when $S(m) > r$ is certain or no longer possible, we have derived the following equation for calculating $CP(p_1, S(m), m)$ exactly:

$$CP(p_1,S(m),m) = \left\{ \begin{array}{ll} 0, & \text{if } m-S(m) > N-r-1 \text{ or} \\ & (m-S(m) > n_1-r_1-1 \text{ and } m \leq n_1) \\ \displaystyle\sum_{j=r-S(m)}^{n_1-m-1} \left[A(j,r_1) \sum_{i=r-S(m)}^{N-(j+m+1)-1} A(i,r) \right], & \text{if } m-S(m) \leq n_1-r_1-1 \text{ and } m \leq n_1 \\ \displaystyle\sum_{i=r-S(m)}^{N-m-1} A(i,r) & \text{if } m-S(m) \leq N-r-1 \text{ and } m > n_1 \\ 1, & \text{if } S(m) > r \end{array} \right\} \tag{2.2}$$

where

$$A(x,y) = \binom{x}{y-S(m)} p_1^{y-S(m)+1} (1-p_1)^{x-\{y-S(m)\}}.$$

The exact CP for the NSC design can also be written recursively as

$$CP(p_1, S(m), m) = \left\{ \begin{array}{ll} 0, & \text{if } m - S(m) > N - r - 1 \\ & \text{or } (m - S(m) > n_1 - r_1 - 1 \text{ and } m \leq n_1) \\ D, & \text{if } m - S(m) \leq N - r - 1 \\ & \text{or } (m - S(m) \leq n_1 - r_1 - 1 \text{ and } m \leq n_1) \\ 1, & \text{if } S(m) > r \end{array} \right\}, \tag{2.3}$$

where

$$D = p_1 CP(p_1, S(m+1), m+1) + (1 - p_1) CP(p_1, S(m), m+1).$$

For a single-stage trial incorporating NSC, the CP can be obtained using these equations by omitting the conditions relating to r_1 and n_1. In its recursive form, it can be seen that the CP at any point in a trial is a function of the CP at points with at least the same number of responses and more participants, among such points that are possible in the trial. By "possible", we simply mean combinations of $S(m)$ and m that may occur given the design parameters of a specific trial.

Consider a grid of CP values for an NSC design based on the number of responses $S(m)$ as rows and the number of participants m as columns, as in Figure 2.1. As an example, let the maximum sample size be $N = 8$ and the rejection boundary be $r = 4$, as in Figure 2.1d. The CP at the point $(S(m) = 3, m = 4)$, that is, $CP(p_1, 3, 4)$, is a function of $CP(p_1, 3, 5)$ and $CP(p_1, 4, 5)$, which in turn are functions of $CP(p_1, 3, 6)$ and $CP(p_1, 4, 6)$, and $CP(p_1, 4, 6)$ and $CP(p_1, 5, 6)$ respectively. As an example, Figure 2.3 is a reproduction of Figure 2.1d with the CP at each point added, for response rate $p_1 = 0.4$.

S(m)	m=1	2	3	4	5	6	7	8
0	0.09	0.03	0					
1	0.28	0.17	0.07	0				
2		0.46	0.32	0.18	0.06	0		
3			0.66	0.52	0.35	0.16	0	
4				0.87	0.78	0.64	0.40	0
5					1	1	1	1
6		Continue						
7		No go decision						
8		Go decision						

Fig. 2.3 Illustrative diagram of NSC designs, including CP at each point, for response rate $p_1 = 0.4$.

2.1.7.1 Accounting for early stopping due to stochastic curtailment

For designs that incorporate NSC, the trial stops and a no go decision is taken if $CP = 0$. The trial stops and a go decision is taken if $CP = 1$. The CP at any point can be obtained using Equation (2.2) directly. For designs that incorporate SC, the trial will additionally end at any point where $0 < CP < \theta_F$ or $\theta_E < CP < 1$, for some specified $\theta_F, \theta_E \in [0,1], \theta_F < \theta_E$. As the CP is a function of later points in the trial, the predetermined decision to end a trial at any point where $0 < CP < \theta_F$ causes the CP of such points to become zero. Conversely, points where $\theta_E < CP < 1$ then have a CP of one. This in turn affects the CP of earlier points in trial. As such, when incorporating SC, it is logical to calculate CP at each point using a recursive equation, one value at a time, starting at the point $(S = r, m = N-1)$, where $CP(p_1, r, N-1) = p_1$ by definition. All "earlier" points in the trial, i.e., points such that $m < N-1$, are either a function of $CP(p_1, r, N-1)$ or are terminal points. For points with more responses or more participants, $CP(p_1, a, N) = 0$ if $a \leq r$, $a \in \mathbb{Z}^{0+}$, and $CP(p_1, a, b) = 1$ for any $a > r$ and any $b \geq a, b \leq N$. Thus for the SC design, the CP at each point can be obtained using the following equation:

$$CP(p_1, S(m), m) = \left\{ \begin{array}{ll} 0, & \text{if } D < \theta_F \text{ or } m - S(m) > N - r - 1 \text{ or} \\ & (m - S(m) > n_1 - r_1 - 1 \text{ and } m \leq n_1) \\ D, & \text{if } \theta_F \leq D \leq \theta_E \text{ and} \\ & m - S(m) > N - r - 1 \text{ or } (m - S(m) > n_1 - r_1 - 1 \text{ and } m \leq n_1) \\ 1, & \text{if } D > \theta_E \text{ or } S(m) > r \end{array} \right\}. \tag{2.4}$$

Similarly to the equations for the NSC design, Equation (2.4) can also be used to obtain the CP for the m-stage design, which is a single-stage design that incorporates SC, by omitting the conditions relating to r_1 and n_1:

$$CP(p_1, S(m), m) = \left\{ \begin{array}{ll} 0, & \text{if } D < \theta_F \text{ or } m - S(m) > N - r - 1 \\ D, & \text{if } \theta_F \leq D \leq \theta_E \text{ and } \{m - S(m) > N - r - 1\} \\ 1, & \text{if } D > \theta_E \text{ or } S(m) > r \end{array} \right\}. \tag{2.5}$$

2.1.8 Constructing stopping boundaries

Once the CP is obtained for each point in the trial, the terminal points $\mathscr{T}$ are found. These points consist of all lower and upper stopping boundaries, which can be obtained as follows:

$$f_m = \begin{cases} \max\left[S(m)\mathbb{I}\left(CP(p_1, S(m), m) = 0\right)\right] & \text{if } \sum_{S(m)=0}^{m} \mathbb{I}\left[CP(p_1, S(m), m) = 0\right] \geq 1 \\ -\infty & \text{otherwise} \end{cases} \tag{2.6}$$

$$e_m = \begin{cases} \min\left[S(m)\mathbb{I}\left(CP(p_1, S(m), m) = 1\right)\right] & \text{if } \sum_{S(m)=0}^{m} \mathbb{I}\left[CP(p_1, S(m), m) = 1\right] \geq 1 \\ \infty & \text{otherwise} \end{cases} \tag{2.7}$$

The vectors of stopping boundaries are then $\mathbf{f} = (f_1, f_2, \ldots, f_N)$, $\mathbf{e} = (e_1, e_2, \ldots, e_N)$. The set of terminal points $\mathscr{T}$ can be considered to be the union of the points in $\mathbf{f}$ and $\mathbf{e}$ that do not equal $\pm\infty$, and their corresponding number of participants:

$$\mathscr{T} = \left\{(f_1, 1), (f_2, 2), \ldots, (f_N, N) : \mathbf{f}^N \in \mathbb{Z}\right\} \cup \left\{(e_1, 1), (e_2, 2), \ldots, (e_N, N) : \mathbf{e}^N \in \mathbb{Z}\right\} \tag{2.8}$$

2.1.9 Choosing thresholds θ_F and θ_E

We seek a set of values from which ordered pairs of θ_F and θ_E will be created and searched over to find optimal or admissible design realisations, for single optimality criteria or weighted multiple optimality criteria respectively. One could use a uniformly distributed set of possible thresholds to some specified degree of coarseness. By choosing a uniform, coarse set of values to search over, the design search can be fast, though some designs with good operating characteristics may be missed. Conversely, undertaking a search over a fine uniform grid may take far longer and still result in missing potential efficient designs. This is because the effect of a chosen threshold θ_F or θ_E on a trial's operating characteristics depends on the CP at each possible point in the trial. Consider for example, the NSC design shown in Figure 2.1d: for each possible combination of participants so far, m, and number of responses, $S(m)$, there exists some CP, $CP(p_1, S(m), m)$. At the points where the trial stops for a go or no go decision, the CP is equal to one or zero respectively, and is strictly between these values at all other points. In trials of this type, the CP values are not uniformly distributed;

instead, most of the mass is close to zero or one. This is shown for three example trials in Figure 2.4.

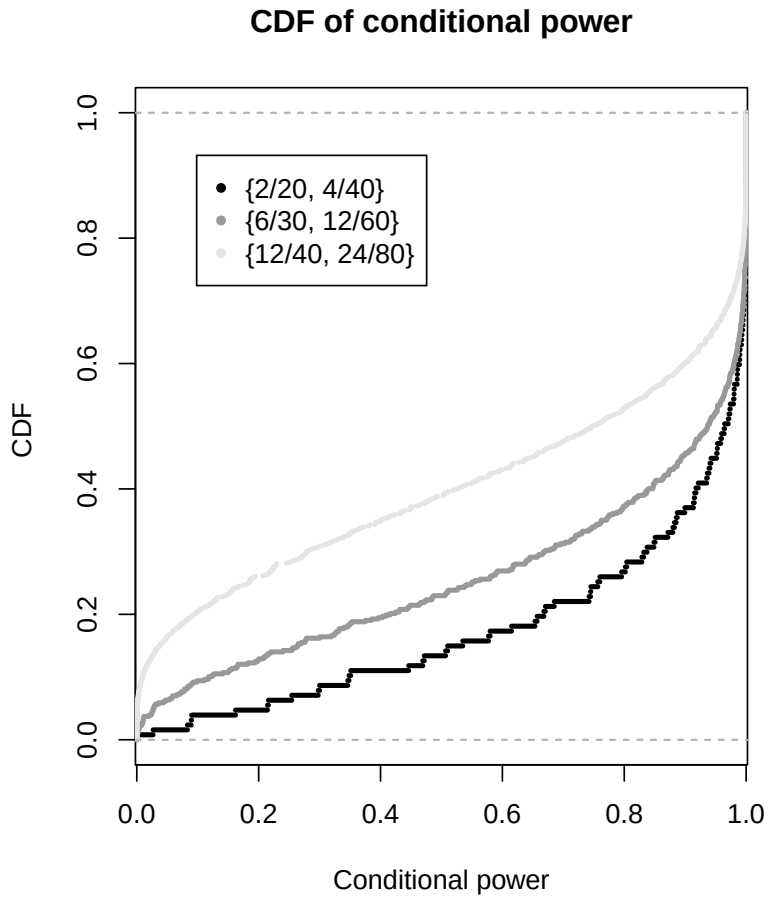


Fig. 2.4 Cumulative distribution function of unique CP values for NSC designs (format $\{r_1/n_1, r/N\}$) with $p_0 = 0.1$ $p_1 = 0.3$, $r = \{Np_0, N(p_0+p_1)/2, Np_1\}$ for $N = \{40, 60, 80\}$ respectively, $r_1 = r/2$, $n_1 = N/2$.

To account for the lack of a uniform distribution, we propose searching over a set of thresholds chosen based on the CP at each point in each possible trial. We obtain every possible value of $CP(p_1, S(m), m)$, including zero and one, for a given combination of $\{r, N\}$ (m-stage) or $\{r_1, n_1, r, N\}$ (SC design). Suppose we allow an upper and lower limit for θ_F and θ_E respectively, termed $\theta_{F_{MAX}}$ and $\theta_{E_{MIN}}$. Then, without loss of generality, a trial-specific

set of thresholds can be defined as

$$\theta = \{CP(p_1, S(m), m) : S(m) = 0, \ldots, r,\ m = 1, \ldots, N : \{r, N\} \in \mathscr{R}, CP \leq \theta_{F_{MAX}} | CP \geq \theta_{E_{MIN}}\}, \tag{2.9}$$

where $\mathscr{R}$ is the family containing all possible sets $\{r, N\}$ (or $\{r_1, n_1, r, N\}$).

2.1.10 Constraining θ

As stated directly above, we allow an upper and lower limit for θ_F and θ_E respectively, termed $\theta_{F_{MAX}}$ and $\theta_{E_{MIN}}$, for example setting $\theta_{E_{MIN}} = 1$ to allow a go decision when CP=1 only. Such limits can be readily incorporated and may be desired, for example, for statistical reasons or to reduce computation time, though our goal is to find the optimal design realisation regardless of θ_E and θ_F values. If a trial using SC reaches $m = N - 1$ participants without a decision being made, then the go or no go decision will depend on the final participant. Specifically, the trial will result in a go decision if the final participant responds and a no go decision if the final participant does not respond. The CP at this point, $CP(p_1, r, N-1)$, is equal to p_1. Under SC, a trial stops for a no go decision if $CP < \theta_F$. However, if the true response rate is great enough to warrant further study, then the probability of a go decision at this point is $p \geq p_1$, and so the trial should not be curtailed for a no go decision at the point $(r, N-1)$. As such, we suggest setting $\theta_{F_{MAX}} = p_1$.

In Section 2.2, optimal design realisations are found for a range of optimality criteria and design parameters. In all such design realisations, for both the SC design and the m-stage design, the upper threshold is greater than 0.97, that is, $\theta_E > 0.97$, despite there being no restriction on θ_E in the design search. This suggests that most, if not all, optimal designs may use an upper threshold in the range $\theta_E \in [0.97, 1.00]$. As such, we suggest a conservative lower bound for the upper threshold of $\theta_{E_{MIN}} = 0.95$.

2.1.10.1 Effect of constraining θ_F and θ_E

For a single-stage design incorporating NSC, the maximum number of possible CP values (as some values may be repeated), including zero and one, is given by

$$|\theta| = (r+1)(N-r)+1,$$

where $|\theta|$ is the cardinality of the set θ and which is a quadratic equation that reaches a maximum at $r = (N-1)/2$. The number of possible CP values increases linearly with N. A'Hern [4] states that the final rejection boundary for a single-stage trial with no curtailment will be approximately

$$r = N(p_0 + [(z_\alpha/(z_\alpha + z_{1-\beta})) \times (p_1 - p_0)]). \tag{2.10}$$

For a single-stage trial with $N = 40$ and design parameters $(\alpha, \beta, p_0, p_1) = (0.05, 0.20, 0.10, 0.30)$, the approximate stopping boundary given by Equation (2.10) is $r = 9.5967$. Setting r equal to the smallest integer greater than this, 10, such a trial would have 331 possible CP values, resulting in 54,615 ordered pairs (θ_F, θ_E) such that $\theta_F < \theta_E$. For a two-stage design incorporating NSC, the number of possible CP values (including zero and one) is given by

$$|\theta| = (r_1+1)(n_1-r_1) + (r-r_1)(N-r)+1.$$

Adding an interim analysis at the midpoint of the example trial $(n_1 = 20)$, with an interim stopping boundary of one half of the example trial final rejection boundary $(r_1 = 5)$, results in 241 possible CP values, from which 28,920 ordered pairs (θ_F, θ_E) such that $\theta_F < \theta_E$ can be created.

Examining the actual CP values for a single-stage trial with design parameters $\alpha = 0.05, \beta = 0.2$, $p_0 = 0.1, p_1 = 0.3$, there are 330 unique values, from which 54,285 ordered

pairs $(\theta_F, \theta_E) : \theta_F < \theta_E$ are possible. Introducing (only) the constraint $\theta_F < p_1$ reduces the the number of possible ordered pairs in this example to 30,414. Further constraining the search to $\theta_E > 0.95$ reduces the number of ordered pairs to 8,325. This is comparable to the number of ordered pairs that would be produced when searching over the uniform sequence $\{0, 0.01, ... 1\}$, which is 5,050. Yet, it should more accurately capture the performance of possible designs.

2.1.11 Controlling maximum length/cardinality of θ

The number of ordered pairs (θ_F, θ_E) within a set θ is $\frac{|\theta|(|\theta|+1)}{2}$. Given how quickly the number of ordered pairs increases with $|\theta|$, we sought to place an upper limit on the cardinality of all such sets. We denote this upper limit by Θ. Once θ has been obtained for a given $\{r,N\}$ or $\{r_1,n_1,r,N\} \in \mathscr{R}$, θ is constrained to contain only CP values less than or equal to $\theta_{F_{MAX}}$ or greater than or equal to $\theta_{E_{MIN}}$. Then $|\theta|$ is checked against Θ. If $|\theta| > \Theta$, its values are placed in order then every other element of θ (excluding zero and one) is removed. This procedure is repeated until $|\theta| \leq \Theta$.

In the design searches for which results are presented later, Θ was set to 10^6, resulting in a maximum number of ordered pairs of approximately 5×10^{11} for each set in $\mathscr{R}$. In the results that follow, the above thinning procedure was used on the set θ prior to applying constraints $\theta \leq \theta_{F_{MAX}}$ or $\theta \geq \theta_{E_{MIN}}$. However, the accompanying code has since been updated, and the thinning procedure is now undertaken after applying the above constraints as described. This means that thinning is less likely to take place, and when it does take place, fewer CP values of interest will be discarded.

2.1.12 Range for final rejection boundary *r*

The computational intensity of searching for admissible designs for all $\mathscr{R}$ increases as $|\mathscr{R}|$ increases. In particular, each possible final rejection boundary r included in a search will

result in additional sets of trials $\{r,N\}$ (or $\{r_1,n_1,r,N\}$) to search over, each with its own set of of CP values, which may considerably increase computational intensity. For example, consider the constraint $r \in [\lfloor Np_0 \rfloor, \lceil Np_1 \rceil]$ for each $N \in [N_{MIN}, N_{MAX}]$, where $[N_{MIN}, N_{MAX}]$ is the range of maximum sample sizes searched over. This constraint is justified directly below, in Section 2.1.13. Taking a typical set of response rates $p_0 = 0.1, p_1 = 0.3$ and range of $N \in [N_{MIN} = 10, N_{MAX} = 50]$, the total number of possible ordered pairs (θ_F, θ_E) for a single-stage design with NSC is 9.13×10^6. Increasing the range of r to $r \in [\lfloor Np_0 \rfloor, \lceil Np_1 \rceil + 1]$ increases the number of possible ordered pairs to 10.84×10^6. This issue is exacerbated in the SC design, as each possible r included in a search will result in the above increase in computation, multiplied by all possible interim design parameters r_1, n_1, $r_1 < \min(r, n_1)$. It is therefore of interest to apply sensible constraints to r.

Unconstrained, the final rejection boundary may take any value $r \in \mathbb{N} < N$. Two approaches to constraining r were investigated: one based on the work of A'Hern [4] and one based on the work of Wald [14]. A'Hern states without proof that for a single-stage design without curtailment, the final rejection boundary r lies in the interval $[Np_0, Np_1]$ [4]; as such, r was constrained to the rounded interval $[\lfloor Np_0 \rfloor, \lceil Np_1 \rceil]$ in the design search. As an alternative, constraining r to the boundaries of Wald's SPRT [14] was also examined. This is a design with no maximum sample size N: the trial simply continues until a stopping boundary is reached. Wald derives lower and upper stopping boundaries for the SPRT to be

$$f_{WALD}(m) = \left(\log \frac{\beta}{1-\alpha} + m \log \frac{1-p_0}{1-p_1} \right) G,$$
$$e_{WALD}(m) = \left(\log \frac{1-\beta}{\alpha} + m \log \frac{1-p_0}{1-p_1} \right) G.$$

where

$$G = \left(\log \frac{p_1}{p_0} - \log \frac{1-p_1}{1-p_0} \right)^{-1},$$

after m participants. The constraint applied r was then $r \in [\lfloor f_{WALD}(N) \rfloor, \lceil e_{WALD}(N) \rceil]$, calculated for each $N \in [N_{MIN}, N_{MAX}]$.

Note that Wald's SPRT results in a lower ESS under $p = p_0$, which we denote $ESS(p_0)$, and ESS under $p = p_1$, which we denote $ESS(p_1)$, than any other design with the same type-I error-rate and power [14, 81]. As such, it is worthwhile to compare how close the ESS of a given design is to the ESS obtained using Wald's SPRT. For Wald's design, the ESS under $p = p_0$ is

$$ESS(p_0) = \frac{(1-\alpha) \log \frac{\beta}{1-\alpha} + \alpha \log \frac{1-\beta}{\alpha}}{p_0 \log \frac{p_1}{p_0} + (1-p_0) \log \frac{1-p_1}{1-p_0}}.$$

The ESS under $p = p_1$ is

$$ESS(p_1) = \frac{\beta \log \frac{\beta}{1-\alpha} + (1-\beta) \log \frac{1-\beta}{\alpha}}{p_1 \log \frac{p_1}{p_0} + (1-p_1) \log \frac{1-p_1}{1-p_0}}.$$

2.1.13 Design search

2.1.13.1 Previous design searches

Our design search is considerably different to that of Ayanlowo and Redden or Kunz and Kieser [12, 13]. In the approach of Kunz and Kieser, the authors obtain the optimal Simon's design, equivalent to a single trial combination $\{r_1, n_1, r, N\}$, and then examine the effect of SC in the form of $\theta_F \in \{0, 0.01, 0.02, \ldots, 1.00\}$, with no θ_E [13]. Ayanlowo and Redden do likewise, but not only for the optimal Simon's design, but also the minimax Simon's design and A'Hern's single-stage design [4, 12]. Ayanlowo and Redden use the thresholds $\theta_F \in \{0.05, 0.10\}$, again with no θ_E.

We do not calculate the type-I error-rate and power of particular combinations $\{r,N\}$ or $\{r_1,n_1,r,N\}$ prior to adding curtailment, as a design that is feasible under SC may not be feasible before the incorporation of SC. Consequently, if such designs were discarded in advance, they therefore would be missed.

As an example, take the design parameters $(\alpha,\beta,p_0,p_1) = (0.05,0.2,0.1,0.4)$. The single-stage design $\{r,N\} = \{4,21\}$ for $(p_0,p_1) = (0.1,0.4)$ has operating characteristics $(\alpha^*,1-\beta^*) = (0.052,0.963)$ (rounded to 3 d.p.), which would not be feasible if we required a type-I error-rate $\alpha \leq 0.05$ and power $1-\beta \geq 0.85$. However, applying SC by using the thresholds $(\theta_F,\theta_E) = (0.31744,0.99190)$ results in the operating characteristics $(\alpha^*,1-\beta^*) = (0.048,0.859)$ (rounded to 3 d.p.), which is feasible and has $ESS(p_0) = 7.5, ESS(p_1) = 7.6$ (rounded to 1 d.p.). In their results, Ayanlowo and Redden and Kunz and Kieser [12, 13] both show that applying SC to an optimal uncurtailed design can result in a decrease in type-I error-rate and power.

2.1.13.2 Proposed design search, in general

For the proposed designs, possible design realisations are found by first setting the desired error-rates α and β, p_0, p_1 and a range for N, $[N_{MIN},N_{MAX}]$. For each $N \in [N_{MIN},N_{MAX}]$ included in the search, a range for r is chosen. The range used in our searches was $[\lfloor Np_0 \rfloor, \lceil Np_1 \rceil]$. However, any range is permitted, and details are provided with regard to this choice in Section 2.1.12.

The sets of $\{r,N\}$ (or $\{r_1,n_1,r,N\}$ in the case of the SC design) are stored as the family of sets $\mathscr{R}$ (Section 2.1.9). For each set in $\mathscr{R}$, the CP of each point in that trial is obtained using Equation (2.4) or (2.5) as appropriate. Once constrained such that all CP values satisfy either $CP \leq \theta_{F_{MAX}}$ or $CP \geq \theta_{E_{MIN}}$, these CPs form the trial-specific set of thresholds θ, that is, $\theta = \{CP(p_1,S(m),m),\, S(m) = 0,\ldots,r,\, m = 1,\ldots,N : (r,N) \in \mathscr{R}, CP \leq \theta_{F_{MAX}} | CP \geq \theta_{E_{MIN}}\}$ (Section 2.1.9). θ may be reduced in size if large (see Section 2.1.11). Within each set

$\{r,N\}$ (or $\{r_1,n_1,r,N\}$), the error-rates α^* and β^*, $ESS(p_0)$ and $ESS(p_1)$ are found for ordered pairs $(\theta_F,\theta_E) \in \theta : \theta_F < \theta_E$. Each $\{r,N,\theta_F,\theta_E\}$ (or $\{r_1,n_1,r,N,\theta_F,\theta_E\}$) describe a particular realisation of a given design, with its own operating characteristics. Among the design realisations obtained, the design realisations that are dominated are discarded. What remains is a collection of admissible designs. The designs that minimise $ESS(p_0)$ and $ESS(p_1)$ respectively are termed the p_0- and p_1-optimal designs. The designs that minimise $ESS(p_0)$ and $ESS(p_1)$ respectively among the subset of designs that minimise N are termed the p_0- and p_1-minimax designs. These terms are analogous to the terms H_0- and H_1-optimal and H_0- and H_1-minimax used by Mander and Thompson [8]. The ESSs of the p_0- and p_1-optimal and p_0- and p_1-minimax admissible designs of the proposed designs will be compared to those of Simon's design, Mander and Thompson and the NSC design, and additionally to those of the designs found using the SPRT of Wald [14] in the case of the p_0- and p_1-optimal criteria.

2.1.14 Design search, in detail

With the caveats in regards to constraining r (Section 2.1.12), $|\theta|$ (Section 2.1.11) and θ_F and θ_E (Section 2.1.9) in place, the search procedure for finding designs can be explained in more detail. This is described in words directly below, and more formally using pseudocode in Algorithm 1. Note that here and in subsequent algorithms and descriptions, some functions are called within a loop for the sake of simplicity, but in the actual code they are vectorised. That is, a function call is simultaneously applied to either every element of a vector or every row or column of a matrix (rather than separately). Other minor aspects of design searches have also been omitted to increase clarity. The design search is as follows:

- Find all sets $\{r,N\}$ (or $\{r_1,n_1,r,N\}$) in $\mathscr{R}$.
- For each set $\{r,N\}$ (or $\{r_1,n_1,r,N\}$) in $\mathscr{R}$,

- Find all CP values.
- Constrain CP values: $CP \leq \theta_{F_{MAX}} | CP \geq \theta_{E_{MIN}}$. This is the set θ.
- If $|\theta| > \Theta$, remove elements as described above, until $|\theta| \leq \Theta$.
- Find all ordered pairs of θ.
- Group the ordered pairs by θ_E and order each group, resulting in one (ordered, ascending) vector of θ_F values for each unique θ_E.
- Sort the unique θ_E values from largest to smallest.
- For each θ_E and corresponding vector of θ_F values:
  - Use the binary search algorithm to find the smallest θ_F that gives a type-I error-rate less than or equal to α. In general, the binary search algorithm is a method that finds a target value within a sorted array: in this case, the target value is the smallest element of the current θ_F vector that results in a type-I error-rate less than α. It is similar to the bisection method, as it bisects the possible values at each iteration. If the type-I error-rate is greater than α, increase θ_F, or else if the type-I error-rate is less than α, decrease θ_F. The bisecting of this vector continues until we find the smallest θ_F that gives a type-I error-rate less than α.
  - From this θ_F and for each subsequent θ_F value in the vector, record the design realisation's type-I error-rate, power, ESS(p_0) and ESS(p_1), as all remaining θ_F values will have the correct type-I error-rate. As θ_F increases, power decreases, therefore stop as soon as power drops below the required power.
- Remove dominated design realisations.

Algorithm 1: Search procedure for a single $\mathscr{R}$

```
for each R do
    k ← 1;
    θ ← call obtainCPvalues(r,N) ;
    θ ← call constrainWRTthetaFmaxthetaEmin(θ, θ_FMAX, θ_EMIN) ;
    while length (θ) > Θ do
        θ ← call halveThetaLength(θ);
    end
    ordered.pairs.matrix ← call findOrderedPairs(θ);
    thetaE.vals ← call unique(ordered.pairs.matrix[,2]);
    thetaE.vals ← call sortDecreasing(thetaE.vals);
    no.thetaE.vals ← call length(thetaE.vals);
    for i = 1 to no.thetaE.vals do        // for each unique θ_E, obtain a corresponding
     vector of θ_F values from the ordered pairs matrix:
        j ← 1;
        for row = 1 to nrow(ordered.pairs.matrix) do
            if ordered.pairs.matrix[row, 2] = thetaE.vals[i] then
                current.thetaF.vec[j] ← ordered.pairs.matrix[row, 1];
                j ← j+1;
            end
        end
        // begin binary search, bisecting the current vector of θ_F values:
           if type-I error-rate > α, increase θ_F, otherwise decrease θ_F.
        a ← 1;
        b ← nrow(current.thetaF.vec);
        d ← ceiling((b-a)/2);
        while (b-a)>1 do
            output ← call findOCs(r,N, θ_F=current.thetaF.vec[d], θ_E=thetaE.vals[i]);
            type.I.err ← call findAlpha(output);
            if type.I.err ≤ α then
                b ← d;
            else
                a ← d;
            end
            d ← a + floor((b-a)/2);
        end
        // We can now proceed moving sequentially from index==b (or break).
        output ← call findOCs(r,N, θ_F=current.thetaF.vec[b], θ_E=thetaE.vals[i]);
        type.I.err ← call findAlpha(output);
        pwr ← call findPower(output);
        if type.I.err ≤ α & pwr ≥ power then
            while pwr ≥ power & b ≤ nrow(current.thetaF.vec) do
                designOCs.matrix[k, ] ← findOCs(r,N, θ_F=current.thetaF.vec[b],
                  θ_E=thetaE.vals[i]);
                pwr ← call findPower(designOCs.matrix[k, ]);
                k ← k+1;
                b ← b+1;
            end
        else
            break
        end
    end
    designOCs.matrix ← call discardDominatedDesigns(designOCs.matrix)
end
```

2.1.15 Comparison of existing and proposed designs (1): design properties

Key differences between existing designs (Simon, Mander and Thompson, NSC, Kunz and Kieser and Ayanlowo and Redden) and our proposed designs are shown in Table 2.1, and in a taxonomy of possible two-stage designs in Table 2.2. In summary, The SC and m-stage designs allow stopping at any point in the trial, not only when the final go decision is certain or not possible, but also likely or unlikely, therefore using SC; a different set of thresholds are examined for each possible trial and exact distributions are used to obtain operating characteristics that are free from simulation error.

	Simon	MT	KK	AR	NSC	SC	m-stage
Allows stopping for go decision	N	Y	N	N	Y	Y	Y
Exact results (i.e., no simulation)	Y	Y	N	N	Y	Y	Y
Allows stopping after each observation	N	N	Y	N	Y	Y	Y
Allows NSC	N	N	Y	Y	Y	Y	Y
Allows SC for no go decision	N	N	Y	Y	N	Y	Y
Allows SC for go decision	N	N	N	N	N	Y	Y
Trial-specific θ's investigated	—	—	N	N	—	Y	Y

Table 2.1 Comparison of methods. MT: Mander and Thompson; KK: Kunz and Kieser; AR: Ayanlowo and Redden.

2.1.16 The loss function

Jung et al. [36] introduced the concept of choosing a design based not on a single optimality criterion, but instead on a combination of two optimality criteria, weighted in importance by an investigator. This was extended by Mander et al. [37] to an expected loss function with weights on three optimality criteria: $ESS(p_0)$, $ESS(p_1)$ and maximum sample size N. The expected loss function is

$$L = w_0 ESS(p_0) + w_1 ESS(p_1) + (1 - w_0 - w_1)N,$$

Allows stopping for go and no go	Allows stopping at any point	Allows NSC	Allows SC	Two-stage design
N	N	N	N	Simon
N	N	N	Y	NA
N	N	Y	N	—
N	N	Y	Y	Simon + AR*†
N	Y	N	N	NA
N	Y	N	Y	NA
N	Y	Y	N	—
N	Y	Y	Y	Simon + KK**†
Y	N	N	N	MT
Y	N	N	Y	NA
Y	N	Y	N	—
Y	N	Y	Y	NSC+AR*†
Y	Y	N	N	NA
Y	Y	N	Y	NA
Y	Y	Y	N	NSC
Y	Y	Y	Y	SC, NSC+KK**†

Table 2.2 Taxonomy of two-stage methods. KK: Kunz and Kieser; AR: Ayanlowo and Redden. MT: Mander and Thompson. *The approach of Ayanlowo and Redden uses $\theta_F \in \{0.05, 0.10\}$. **The approach of Kunz and Kieser uses $\theta_F \in \{0, 0.01, \ldots, 1\}$. † SC for no go decision only. NA: Design not possible. Dash: Design possible.

where $w_0, w_1 \in [0,1]$ and $w_0 + w_1 \leq 1$. In Mander et al., the admissible design, previously defined as the design realisation with the smallest expected loss for a given set of weights, was plotted on a grid of all possible combinations of weights. We extend this concept to allow the comparison of design realisations across differing design approaches: for each combination of weights, the design realisation with the lowest loss, L, across all design approaches (Simon, Mander and Thompson, SC, m-stage, etc.) is found, and this design realisation is termed the *omni-admissible* design realisation for that combination of weights. Across a grid of possible combinations of weights, the design approach to which each omni-admissible design belongs is plotted. In addition, the difference between the expected loss of admissible design realisations at each set of weights is quantified, for certain pairs of design types. The values have no intrinsic meaning; their only purpose is to facilitate the comparison of designs, with a small difference indicating that the compared design realisations perform similarly. For brevity, the admissible design realisations for each design are plotted for the first scenario only. The remainder are given in the Appendix. From these plots, the number of admissible design realisations, and the range of weights for which each admissible design realisation has the lowest loss, can be seen.

2.1.17 Inference: estimation of response rate

The most important aspect of a phase II trial is to decide if a treatment is worth further study. However, it is also important to undertake inference using the trial data, to help make decisions about possible future trials. In particular, one may estimate the response rate of the treatment. The MLE of the response rate is the observed response rate, $\hat{p} = S(m)/m$, where m represents the number of participants after which the trial stopped. This estimator is biased in trials that allow stopping at an interim analysis, and this may be a source of concern for investigators with a strong interest in point estimation. However, there are a range of estimators available that aim to reduce this bias. In general, such estimators have only been

previously presented for two-stage designs, with the notable exceptions of Girshick et al. [82] and Jung and Kim [83]. To address possible concerns regarding point estimation in curtailed designs, we therefore examine estimates of the response rate across existing and novel designs, for five estimators, extended here to the multi-stage case: the naïve estimator, that is, the MLE above; the bias-adjusted estimator [84]; the simplified bias subtraction estimator [85]; the median unbiased estimator (MUE) [86] and the uniformly minimum-variance unbiased estimator (UMVUE) [83]. A range of estimators are considered as there is no single estimator that performs best in all situations. We evalaute the bias, $Bias(\hat{p}|p) = E(\hat{p}|p) - p$, and the root mean square error (RMSE). The RMSE, $RMSE(\hat{p}|p) = \sqrt{(Bias(\hat{p}|p))^2 + Var(\hat{p}|p)}$, where $Var(\hat{p}|p) = E(\hat{p}^2|p) - E(\hat{p}|p)^2$, is equivalent to taking the square root of the weighted average of the squared distances between each possible point estimate and the true value. The results are shown for the p_0-optimal admissible designs of each design approach.

2.1.17.1 Point estimators for multi-stage trials

The bias-subtracted and bias-adjusted estimators are described in terms of the expected value of the response rate and its bias. The expected estimate of the response rate, $\hat{p}$, can be obtained by taking the product of the observed response rate for each possible terminal point and its probability given some true p, and summing across all possible terminal points:

$$\mathbb{E}(\hat{p}|p,\mathbf{e},\mathbf{f},\mathbf{n}) = \sum_{j=1}^{J} \sum_{S(n_j)=0}^{n_j} \mathbb{I}[\{S(n_j),n_j\} \in \mathscr{T}]\hat{p}(S(n_j),n_j)U(S(n_j),n_j|p,\mathbf{e},\mathbf{f},\mathbf{n}),$$

where $\hat{p}(S(n_j),n_j)$ is the observed response rate $\hat{p}$ at the point $(S(n_j),n_j)$, $\mathbf{e} = (e_1,e_2,\ldots,e_J)$ and $\mathbf{f} = (f_1,f_2,\ldots,f_J)$ are the vectors of stopping boundaries for go and no go decisions respectively at each stage, and $\mathbf{n} = (n_1,n_2,\ldots,n_J)$ is the vector of sample sizes in each stage. For continuous monitoring, $n_1 = n_2 = \cdots = n_J = 1$. The bias, variance

and RMSE are as follows:

$$\begin{aligned}
\text{Bias}(\hat{p}|p,\mathbf{e},\mathbf{f},\mathbf{n}) &= \mathbb{E}(\hat{p}|p,\mathbf{e},\mathbf{f},\mathbf{n}) - p \\
\text{Var}(\hat{p}|p,\mathbf{e},\mathbf{f},\mathbf{n}) &= \mathbb{E}(\hat{p}^2|p,\mathbf{e},\mathbf{f},\mathbf{n}) - \mathbb{E}(\hat{p}|p,\mathbf{e},\mathbf{f},\mathbf{n})^2 \\
\text{RMSE}(\hat{p}|p,\mathbf{e},\mathbf{f},\mathbf{n}) &= \sqrt{\text{Bias}(\hat{p}|p,\mathbf{e},\mathbf{f},\mathbf{n})^2 + \text{Var}(\hat{p}|p,\mathbf{e},\mathbf{f},\mathbf{n})}
\end{aligned}$$

As stated above, the naïve estimator for p is simply $\hat{p}_{naive} = S(m)/m$. The bias-subtracted estimator is then

$$\hat{p}_{bias-sub} = \hat{p}_{naive} - \text{Bias}(\hat{p}_{naive}|\hat{p}_{naive},\mathbf{e},\mathbf{f},\mathbf{n})$$

The bias-adjusted estimator is the numerical solution to

$$\hat{p}_{bias-adj} = \hat{p}_{naive} - \text{Bias}(\hat{p}_{naive}|\hat{p}_{bias-adj},\mathbf{e},\mathbf{f},\mathbf{n})$$

The median unbiased estimator, $\hat{p}_{MUE}$ is obtained by numerically searching for the value of p that would make the p-value equal to 0.5:

$$\text{p-val}(S(m),m|\hat{p}_{MUE}) = 0.5,$$

where the p-value is computed as the sum of the probability of possible outcomes with a larger value of the UMVUE. The UMVUE for a single-arm multi-stage binomial outcome trial was derived by Jung and Kim [83]. At some point $(S(m),m)$, the UMVUE is

$$\hat{p}_{UMVUE} = \mathbb{E}(\hat{p}^{(m_1)}|S(m),m),$$

the expected value of the response rate after some m_1 participants, denoting the first point at which a decision may be made. That is, $m_1 = \min(i) : e_i \neq \infty \vee f_i \neq -\infty$. The estimates for all estimators are obtained using the R package *singlearm* [87].

2.1.18 Less frequent monitoring

Continuous monitoring, that is, undertaking an interim analysis after every participant, maximises the potential benefit of using SC. However, continuous monitoring may not be possible in practice. This may be because the trial recruitment rate is expected to be high, because of the manner in which the results are expected to be reported, or for some other reason. In such instances, less frequent monitoring may be planned. This can be described as sequential monitoring. We describe a design approach for SC using sequential monitoring in terms of specified block sizes B, though it is also possible to specify the number of stages instead. An interim analysis is undertaken after every block of B participants, at which point the number of responses is compared to corresponding lower and upper stopping boundaries. If the number of responses is less than the lower boundary, the trial stops for a no go decision. If it exceeds the upper boundary, the trial stops for a go decision. Otherwise, the trial continues.

The recursive equation used to calculate CP, Equation (2.4), may still be used, with D now generalised to handle blocks of size B:

$$D(B) = \sum_{i=0}^{B} p_1^i (1-p_1)^{B-i} CP(p_1, S(m+i), m+B).$$

With this generalisation, Equation (2.4) can now be used to obtain CP for every possible number of responses for $n \in \{B, 2B, \ldots, N\}$ participants, from which lower and upper stopping boundaries can be obtained. The idea is that lower and upper stopping boundaries exist only at these interim analyses. The resulting design search proceeds in the same manner as for continuous monitoring, but only recording the CP values at the interim analyses. A consequence of this is that the number of CP values $|\theta|$ is reduced for each $\{r, N\}$. The number of possible maximum sample sizes to search over is also reduced, to

$N \in \{N_{MIN}, N_{MIN}+B, N_{MIN}+2B, \ldots, N_{MAX}\}$. As a result, the computational intensity of the design search is greatly reduced.

2.1.19 Comparison of existing and proposed designs (2): results

In Section 2.2 we present one real data example and three scenarios, comparing the proposed designs to existing designs in a variety of ways.

Using the real data example, we compare designs in terms of $\text{ESS}(p_0)$, $\text{ESS}(p_1)$ and N. Also using this example, Simon's design is compared to the m-stage design in terms of median sample size and other quantiles, and for both continuous and less frequent monitoring, in the manner described in Section 2.1.6.

For the m-stage design, we compare the final rejection boundaries of the admissible design realisations to the ranges created by following the equations of A'Hern and Wald [4, 14].

The design parameters for the three considered scenarios are identical to those used by Jung et al. [36]. For all scenarios, existing designs are compared to the proposed designs by finding the design realisations for each design type that satisfy each of the single optimality criteria p_0-optimal, p_1-optimal, p_0-minimax and p_1-minimax. In order to compare designs across multiple criteria simultaneously, $\text{ESS}(p_0)$, $\text{ESS}(p_1)$ and N are combined using the loss function of Mander et al. [37], which assigns a weight to each criterion. We compare the admissible design realisations of each design type across a grid of possible combinations of weights, and produce plots showing the design approach that contains the omni-admissible design realisation for each combination of weights.

The effect of reducing monitoring frequency is examined, allowing flexibility between a single interim analysis and continuous monitoring.

We compare design realisations in terms of estimates of response rate, for a range of estimators.

2.2 Results

2.2.1 Real data example

Kunz and Kieser [13] present a real data example from Sharma et al. [79]. In this trial, the following design parameters were chosen: $\alpha = 0.05, \beta = 0.1, p_0 = 0.2, p_1 = 0.4$. Kunz and Kieser compare the following combinations of designs to the (p_0-)optimal Simon design, with SC permitted for a no go decision only:

- "Simon + AR": NSC for no go only, SC in stage 2 only;
- "Simon + KK": NSC for no go only, SC in both stages;
- "CC + AR": NSC for go and no go, SC in stage 2 only;
- "CC + KK": NSC for go and no go, SC in both stages,

where AR is the design of Ayanlowo and Redden, KK is the design of Kunz and Kieser and CC is the design of Chi and Chen. The results for threshold $\theta_F = 0.4$ from Kunz and Kieser [13] are reported here, as the authors report $ESS(p_0)$ for only $\theta_F = 0.4$ and $\theta_F = 0.6$ and state that trials using $\theta_F = 0.6$ do not achieve adequate power. Table 2.3 contains the operating characteristics for these designs to as great an extent as possible, and additionally:

- Simon: p_0-optimal Simon's design;
- CC: The NSC design of Chi and Chen;
- SC_1: a realisation of the SC design chosen for its resemblance to the other compared trials in terms of maximum sample size N;
- SC_2: the p_0-optimal SC design;
- m-stage$_1$: a realisation of the m-stage design chosen for its resemblance to the other compared trials in terms of maximum sample size N;

- m-stage$_2$: the p_0-optimal m-stage design;
- Wald: Wald's SPRT.

The operating characteristics of Simon+AR, Simon+KK, CC+AR and CC+KK were obtained from the results of Kunz and Kieser [13] and from `Stata` using the `simontwostage` package [88]. The maximum N searched over for the SC and m-stage designs respectively is $N = 58$ and $N = 94$, due to computational intensity, with the range of r chosen based on the boundaries of Wald's SPRT [14].

Design	r_1	n_1	r	N	α^*	$1-\beta^*$	$ESS(p_0)$	$ESS(p_1)$	θ_F	θ_E
Simon	4	19	15	54	0.048	0.904	30.4	51.6	–	–
CC	4	19	15	54	0.048	0.904	28.2	37.6	0.000	1.000
Simon + AR	4	19	15	54	–	0.882*	26.6	–	0.400	1.000
Simon + KK	4	19	15	54	0.038	0.857*	21.2	–	0.400	1.000
CC + AR	4	19	15	54	–	0.882*	25.4	–	0.400	1.000
CC + KK	4	19	15	54	–	0.857*	21.0	–	0.400	1.000
SC_1	2	14	15	54	0.050	0.901	23.0	26.6	0.164	0.998
SC_2	4	21	16	58	0.050	0.900	22.6	25.5	0.199	0.998
m-stage$_1$	–	–	15	52	0.049	0.909	25.3	25.8	0.135	0.996
m-stage$_2$	–	–	26	94	0.049	0.902	22.1	23.3	0.228	0.998
Wald	–	–	–	–	0.050	0.900	21.8	22.7	–	–

Table 2.3 Comparison of designs, with design parameters $(\alpha, \beta, p_0, p_1) = (0.05, 0.10, 0.20, 0.40)$. CC: Chi and Chen. AR: Ayanloyo and Redden. KK: Kunz and Kieser. Blanks in α^* and $ESS(p_1)$ due to data not being included in Kunz and Kieser and not being reproducible using the Stata package `simontwostage`. *Median values, from simulation.

It can be seen from Table 2.3 that with the exception of Wald, the designs with the lowest $ESS(p_0)$ are Simon+KK and CC+KK, which use a threshold of $\theta_F = 0.4$ and allow stopping at any point. However, these designs both have power $1-\beta^* = 0.857 < 1-\beta = 0.9$. The designs Simon+AR and CC+AR also have power less than $1-\beta = 0.9$. This is due to the nature of the design search, whereby an optimal (or minimax) Simon design is obtained that satisfies some $(\alpha, 1-\beta)$ requirement, then some form of curtailment is applied, which

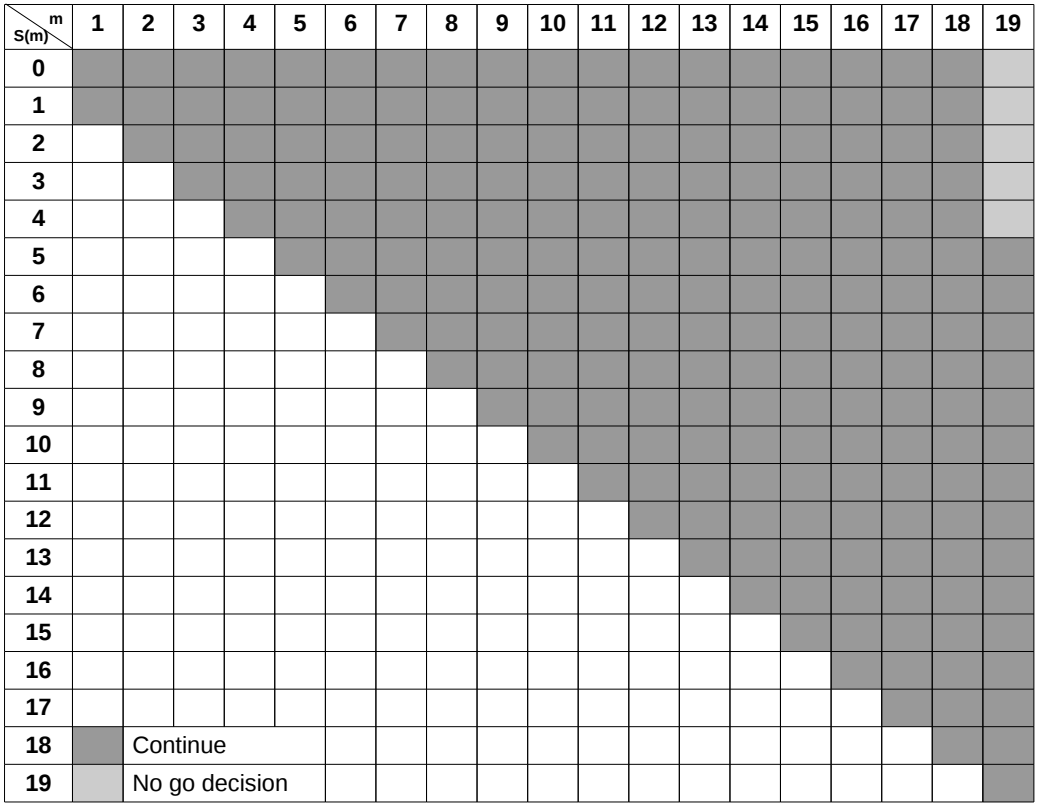
decreases both the type-I error-rate and power when the curtailment is stochastic and for a go no decision only.

The four design realisations obtained using an approach from one of the two proposed designs achieve a lower $ESS(p_0)$ than all feasible design realisations with the exception of Wald, while achieving the necessary type-I error-rate and power. They also have lower thresholds for stopping for a no go decision compared to other designs that use SC, with a maximum of $\theta_F = 0.228$ compared to $\theta_F = 0.4$. Furthermore, the first m-stage design has a lower maximum sample size than all other designs.

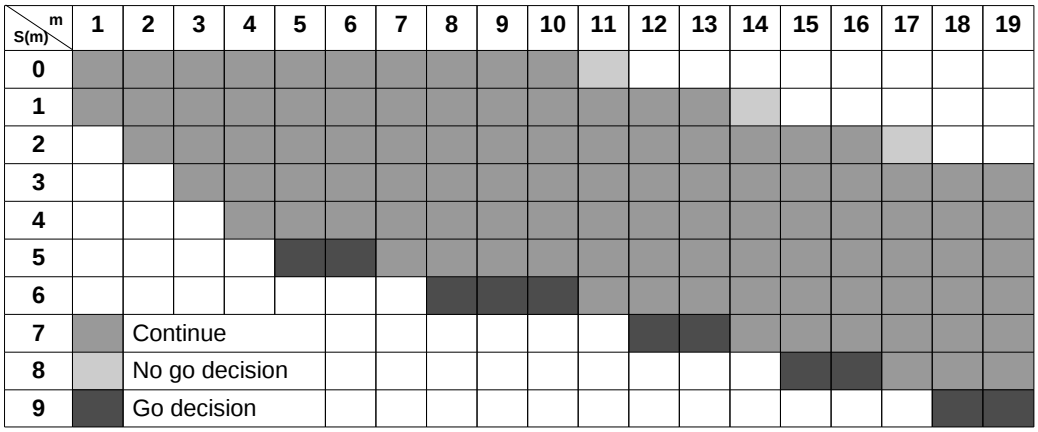
The study by Sharma et al. [79] ended at the first stage, with zero responses out of 19 participants. Using NSC only, the study would have ended after 15 participants. However, using the m-stage design optimised for $ESS(p_0)$, m-stage$_2$ in Table 2.3, the study would have ended after 8 participants. Under the design m-stage$_1$ in Table 2.3, the study would have ended after 11 participants. The latter result is shows in Figure 2.5, which shows go and no go decision boundaries for Simon's optimal design (Figure 2.5a) and the first m-stage design (Figure 2.5b) in Table 2.3 (m-stage$_1$). These figures show all possible decisions that may be made within the first 19 participant responses, which represents the first stage of Simon's design. Simon's design was used in the trial of Sharma et al. [79], while the m-stage design realisation is an example of a design that uses SC and satisfies the required type-I error-rate and power.

2.2.2 Example trials: three scenarios

Three sets of design parameters, or scenarios, were used to compare five design approaches: Simon's design; Mander and Thompson's design; the NSC design; the SC design and the m-stage design. For each scenario and design type, optimal design realisations were obtained that satisfy each of four single optimality criteria. Also for each scenario and design type, a set of admissible design realisations were obtained with regard to the loss function specified



(a) Simon design used in the trial of Sharma et al.[79]: $r_1 = 4, n_1 = 19, r = 15, N = 54$, first stage only.



(b) m-stage design with properties $r = 15, N = 52, \theta_F = 0.135, \theta_E = 0.996$, first 19 participants only

Fig. 2.5 Visualisation of two design realisations satisfying the design parameters of Sharma et al. $(\alpha = 0.05, /beta = 0.1, p_0 = 0.2, p_1 = 0.4)$.

by Mander et al. [37] (Section 2.1.16). For Simon's and Mander and Thompson's design, the maximum sample size searched over was 20% greater than the maximum sample size of the p_0-optimal design, as in Mander et al. [37]. For the NSC and m-stage designs, the maximum sample size searched over was set to 80, approximately 2-3 times greater than the maximum sample size for the optimal Simon design's under the p_0-optimal and p_0-minimax criteria. For the SC design, the maximum sample size was 43 to 47 depending on the scenario, due to computational intensity. For the proposed designs, the range of r was chosen based on the bounds of A'Hern [4] $[\lfloor Np_0 \rfloor, \lceil Np_1 \rceil]$, though the final sets of admissible designs contained only values of r that were also within the (generally stricter) bounds of Wald's SPRT [14]. This is discussed in Section 2.2.3.3. Also reported is $ESS(p_0)$ and $ESS(p_1)$ from Wald's SPRT. As Wald's SPRT seeks to minimise ESS and has no maximum sample size, the ESSs from this test will be compared to those obtained under the p_0- and p_1-optimality criteria.

2.2.3 Scenario 1: design parameters $(\alpha, \beta, p_0, p_1) = (0.05, 0.15, 0.1, 0.3)$

Table 2.4 shows the optimal design realisation for each design approach, for four optimality criteria: p_0-optimal, p_1-optimal, p_0-minimax and p_1-minimax. For all four optimality criteria, the optimal design realisations of the proposed designs outperform those of the existing designs, and use thresholds of $\theta_F < 0.23$ and $\theta_E > 0.98$ in each case. The ESSs using Wald's SPRT are $ESS(p_0) = 13.9, ESS(p_1) = 13.9$, comparable to those of the m-stage design, $ESS(p_0) = 14.1$ under p_0-optimality, $ESS(p_0) = 14.3$ under p_1-optimality and $ESS(p_1) = 14.4$ under both p_0- and p_1-optimality. Note that for this scenario, under both p_0- and p_1-minimax criteria, the SC design realisation happens to be the m-stage design realisation with the addition of an interim analysis. Furthermore, all five design types have p_0- and p_1-minimax design realisations with the same maximum sample size and final rejection boundary $(r = 5, N = 27)$. As such, all differences in operating characteristics between these

design realisations are due to the existence/choice of interim analysis, corresponding interim stopping boundary and the use of curtailment (stochastic, non-stochastic or neither).

	r_1	e_1	n_1	r	N	$ESS(p_0)$	$\%S_{p_0}$	$ESS(p_1)$	$\%S_{p_1}$	θ_F	θ_E
p_0-optimal											
Simon	1	–	11	6	35	18.3	1.00	32.3	1.00	–	–
MT	1	4	11	6	35	18.2	1.00	27.2	0.84	–	–
NSC	1	–	13	5	28	17.6	0.97	18.5	0.57	0.000	1.000
SC	4	–	27	7	41	14.3	0.78	15.0	0.46	0.186	0.993
m-stage	–	–	–	13	80	14.1	0.77	14.4	0.45	0.226	0.997
p_1-optimal											
Simon	2	–	18	5	27	20.4	1.00	26.5	1.00	–	–
MT	0	3	13	6	30	25.1	1.23	20.0	0.76	–	–
NSC	1	–	13	5	28	17.6	0.87	18.5	0.70	0.000	1.000
SC	4	–	24	8	43	15.5	0.76	14.6	0.55	0.126	0.984
m-stage	–	–	–	12	66	14.3	0.70	14.4	0.54	0.189	0.990
Wald's SPRT	–	–	–	–	–	13.9	0.68	13.9	0.52	–	–
p_0-minimax											
Simon	2	–	18	5	27	20.4	1.00	26.5	1.00	–	–
MT	1	4	14	5	27	19.3	0.95	21.0	0.79	–	–
NSC	2	–	18	5	27	19.3	0.95	18.7	0.71	0.000	1.000
SC	0	–	10	5	27	17.1	0.84	16.3	0.62	0.070	0.990
m-stage	–	–	–	5	27	18.7	0.92	16.6	0.63	0.084	0.990
p_1-minimax											
Simon	2	–	18	5	27	20.4	1.00	26.5	1.00	–	–
MT	1	4	15	5	27	20.3	0.99	20.8	0.78	–	–
NSC	2	–	18	5	27	19.3	0.95	18.7	0.71	0.000	1.000
SC	4	–	24	5	27	18.8	0.92	15.8	0.60	0.050	0.986
m-stage	–	–	–	5	27	18.7	0.92	16.6	0.63	0.084	0.990

Table 2.4 Optimal design realisations for each design type, Scenario 1: $(\alpha, \beta, p_0, p_1) = (0.05, 0.15, 0.10, 0.30)$. For all designs, the requisite type-I error-rate and power is achieved. Columns $\%S_{p_0}$ and $\%S_{p_1}$ show ESS as a proportion of Simon's design under $p = p_0$ and $p = p_1$ respectively. MT: Mander and Thompson.

Figure 2.6 (left) shows the design approach to which the omni-admissible design realisation belongs, for each combination of weights. The omni-admissible design belongs to either the SC design or the *m*-stage design. The difference in expected loss between the SC and *m*-stage admissible design realisations for each combination of weights is shown in

Figure 2.6 (right). It shows that the admissible m-stage design realisations have a slightly lower loss score than those of the SC design realisations near the triangle's hypotenuse, that is, when there is low weight on maximum sample size N. For much of the surface of weight combinations, the difference in loss score is in favour of the SC design but negligible, including where both w_0 and w_1 are close to zero and the weight of N is close to one. The maximum difference in loss score between a superior SC design and inferior m-stage design is 3.0. The range of loss score across all admissible design realisations of all design types is (14.1, 79.4), with median 23.1 (IQR [19.2, 26.6]).

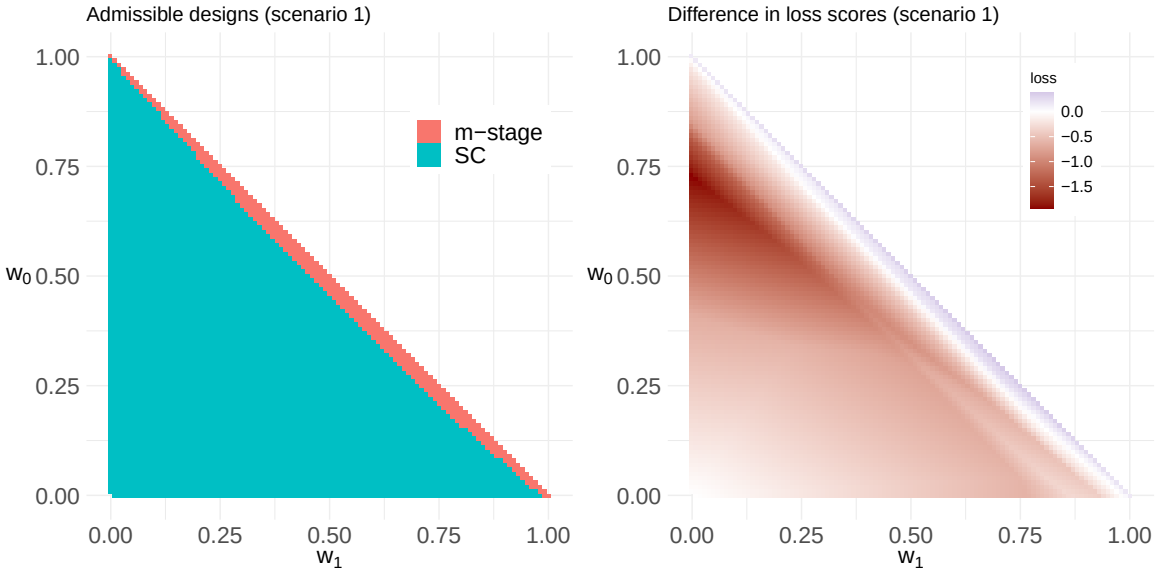


Fig. 2.6 Type of design to which the omni-admissible design realisation belongs and difference in loss scores between the SC and m-stage admissible design realisations (positive favours m-stage), scenario 1 $(\alpha, \beta, p_0, p_1) = (0.05, 0.15, 0.10, 0.30)$.

2.2.3.1 Admissible design realisations by design type, scenario 1 only

In Figure 2.7, the scenario 1 admissible design realisations are shown for each design type and combination of weights. For completeness, corresponding figures are shown for scenarios 2 and 3 in the Appendix. The plots of admissible design realisations for the Simon and Mander and Thompson's designs, Figure 2.7 (top), match those obtained by Mander et al.,

as do the corresponding plots in the Appendix [37]. The overall results are similar across all three scenarios: the proposed designs generally contain a greater number of admissible design realisations across the combinations of weights examined than the existing designs. This is expected as including SC thresholds necessarily results in an increased number of possible design realisations. For the proposed designs, the admissible design regions often contain slopes parallel to the hypotenuse, suggesting that the admissible design may be more dependent on the weight of N than $ESS(p_0)$ or $ESS(p_1)$ separately. In some cases, this is manifested in long, thin regions near the hypotenuse. At the hypotenuse, where the weight of maximum sample size is zero, the admissible design realisations have the greatest maximum sample size of all admissible design realisations. Conversely, maximum sample size decreases as the weight of N increases (that is, towards the bottom left corners), as would be expected. When the weight of N is not close to one, the proposed designs often have a maximum sample size similar to those that do not employ curtailment.

2.2.3.2 Expected loss, scenario 1 only

Heat maps of expected loss for the admissible design realisations of each design type are shown in Figure 2.8 for scenario 1. The proposed designs have a lower expected loss in general. The proposed designs seem most superior in regions where N is weighted close to zero, and where w_0 is close to zero (that is, where the weight of $ESS(p_0) \approx 0$). Again, analogous plots for scenarios 2 and 3 are provided in the Appendix (Figures A.1 and A.2).

To give more context to the loss function values, more expected loss values are given in Table 2.5. Here, we show the expected loss, by component, for admissible design realisations of each design type for a selection of weights. Some of these weights correspond to the single optimality criteria used in Table 2.4. We have already covered which design types can be considered superior for each combination of weights through Figure 2.6 and for single optimality criteria in Table 2.4, and so here we focus on differences in loss

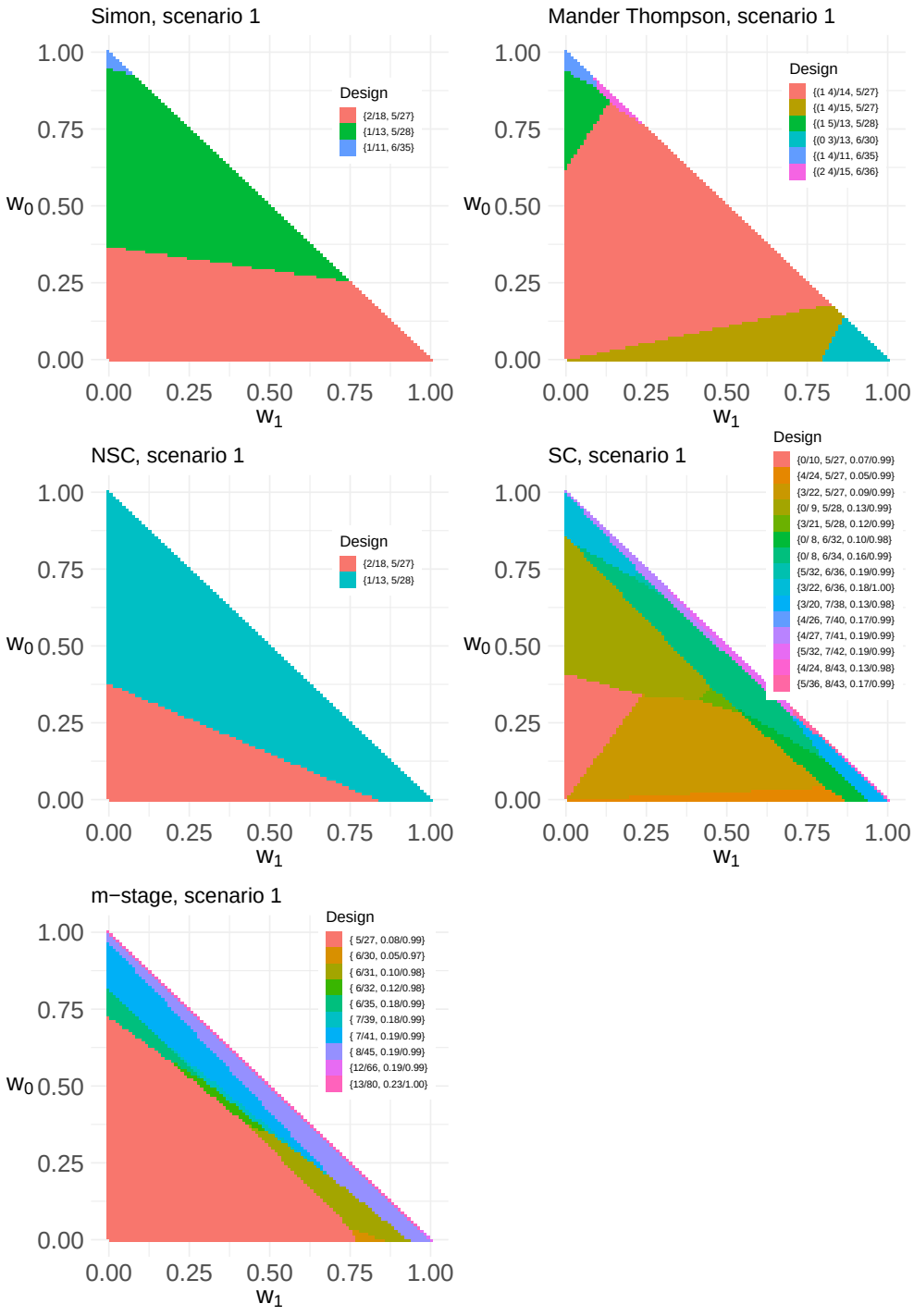


Fig. 2.7 Admissible design realisations for scenario 1 $(\alpha, \beta, p_0, p_1) = (0.05, 0.15, 0.10, 0.30)$. Format of design realisations: Simon, NSC: $\{r_1/n_1, r/N\}$; Mander and Thompson: $\{(r_1\ e_1)/n_1, r/N\}$; SC: $\{r_1/n_1, r/N, \theta_F/\theta_E\}$; m-stage: $\{r/N, \theta_F/\theta_E\}$

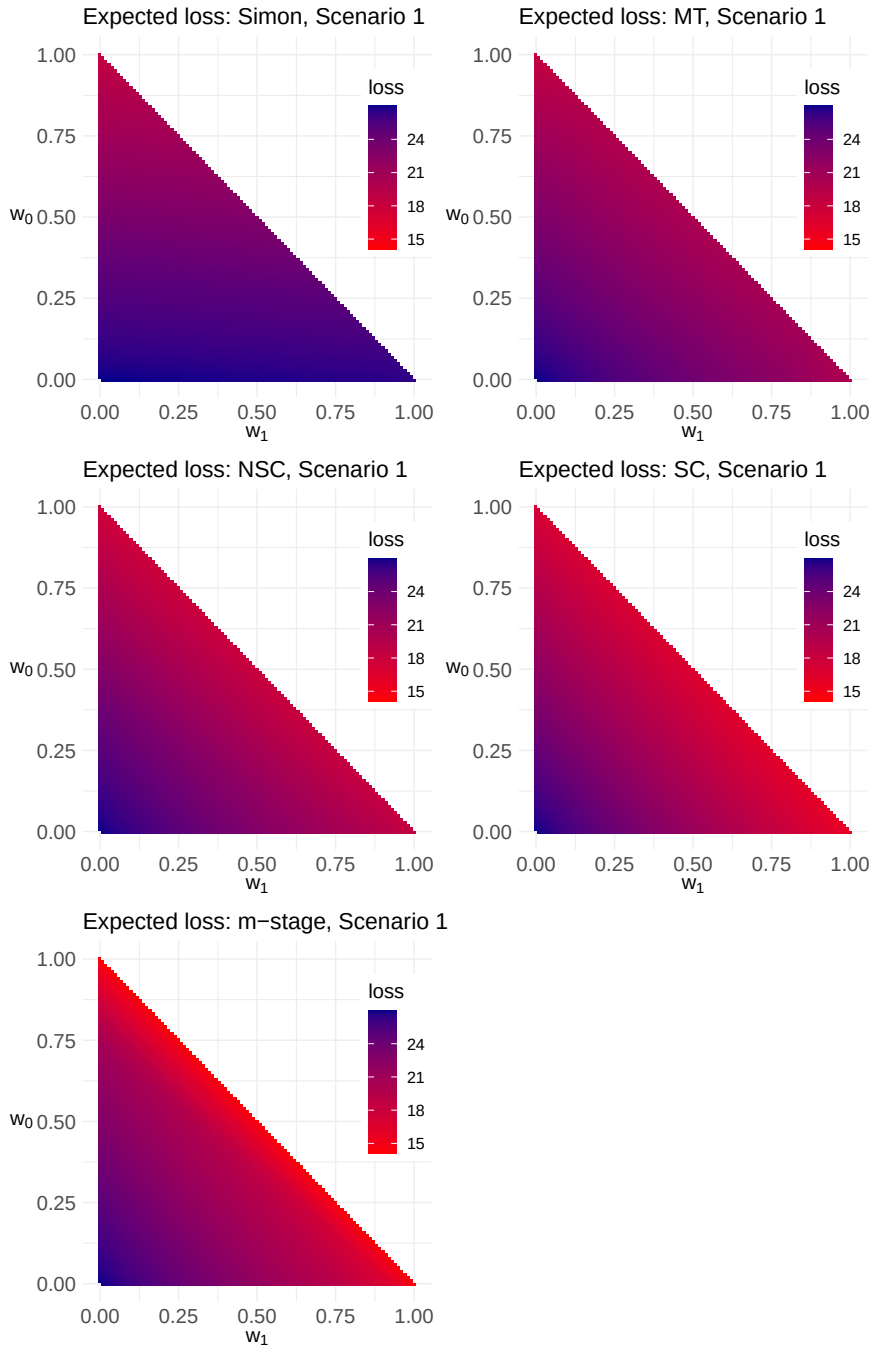


Fig. 2.8 Expected loss for obtained admissible design realisations of each design type, for scenario 1 $(\alpha, \beta, p_0, p_1) = (0.05, 0.15, 0.10, 0.30)$.

scores. The first set of weights, $(w_0 = 1, w_1 = 0)$, corresponds to p_0-optimality, as it focuses only on minimising $ESS(p_0)$. There is little difference between the NSC design and the uncurtailed designs, while the two proposed designs have similar loss scores. The second set of weights, $(w_0 = 0, w_0 = 1)$, corresponds to p_1-optimality, as it focuses only on minimising $ESS(p_1)$. Again the two proposed designs have similar results, while the Mander and Thompson and NSC designs considerably outperform Simon's design. The third set of weights, $(w_0 = 0.01, w_0 = 0)$, corresponds to p_0-minimax, as it essentially focuses on minimising N $(1 - w_0 - w_1 = 0.99)$ while allowing ties to be broken by using a nominal weight of $w_0 = 0.01$ on $ESS(p_0)$. The discrete nature of maximum sample size means that, as all design realisations have the same minimum maximum sample size, the differences between the loss scores are entirely due to $ESS(p_0)$. The weighted differences between the design realisations in terms of $ESS(p_0)$ are small, with the total loss scores identical after rounding to 1 d.p.. The final set of weights, $(w_0 = 1/3, w_1 = 1/3)$, places equal weight on $ESS(p_0)$, $ESS(p_1)$ and N. The two proposed design types have similar (but superior) loss scores to the NSC design, which in turn has a similar loss score to the Mander and Thompson design. Simon's design performs relatively poorly compared to the proposed designs.

2.2.3.3 Comparison of boundaries used by Wald and A'Hern to final rejection boundaries of *m*-stage designs

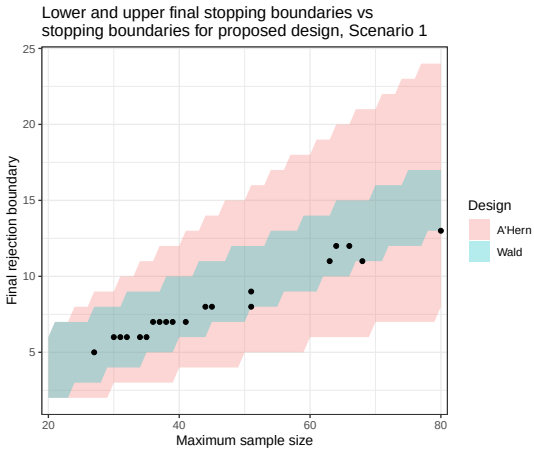
Each design search was undertaken for a range of maximum sample sizes $N \in [N_{MIN}, N_{MAX}]$. A number of admissible design realisations were obtained during each search. This is the set of design realisations for which the expected loss was obtained. For each $N \in [N_{MIN}, N_{MAX}]$ in the design search, the final rejection boundary r was constrained in order to decrease computation time. We chose $r \in [\lfloor Np_0 \rfloor, \lceil Np_1 \rceil]$, as A'Hern [4] states that for single-stage designs, the final rejection boundary must lie within the interval $[Np_0, Np_1]$. Therefore each N in the design search was accompanied by a corresponding interval of final stopping

	$ESS(p_0)$	$w_0ESS(p_0)$	$ESS(p_1)$	$w_1ESS(p_1)$	N	$(1-w_0-w_1)N$	E(L)	$E(L)-\min(E(L))$
$(\mathbf{w_0=1, w_1=0})$								
Simon	18.3	18.3	32.3	0.0	35	0.0	18.3	4.1
MT	18.2	18.2	27.2	0.0	35	0.0	18.2	4.1
NSC	17.6	17.6	18.5	0.0	28	0.0	17.6	3.5
SC	14.3	14.3	15.0	0.0	41	0.0	14.3	0.2
m-stage	14.1	14.1	14.4	0.0	80	0.0	14.1	0.0
$(\mathbf{w_0=0, w_1=1})$								
Simon	20.4	0.0	26.5	26.5	27	0.0	26.5	12.1
MT	25.1	0.0	20.0	20.0	30	0.0	20.0	5.6
NSC	17.6	0.0	18.5	18.5	28	0.0	18.5	4.1
SC	15.5	0.0	14.6	14.6	43	0.0	14.6	0.2
m-stage	14.3	0.0	14.4	14.4	66	0.0	14.4	0.0
$(\mathbf{w_0=0.01, w_1=0})$								
Simon	20.4	0.2	26.5	0.0	27	26.7	26.9	<0.1
MT	19.3	0.2	21.0	0.0	27	26.7	26.9	<0.1
NSC	19.3	0.2	18.7	0.0	27	26.7	26.9	<0.1
SC	17.1	0.2	16.3	0.0	27	26.7	26.9	0.0
m-stage	18.7	0.2	16.6	0.0	27	26.7	26.9	<0.1
$(\mathbf{w_0=1/3, w_1=1/3})$								
Simon	18.7	6.2	27.0	9.0	28	9.3	24.6	4.5
MT	19.3	6.4	21.0	7.0	27	9.0	22.4	2.3
NSC	17.6	5.9	18.5	6.2	28	9.3	21.4	1.3
SC	15.7	5.2	16.6	5.5	28	9.3	20.1	0.0
m-stage	18.7	6.2	16.6	5.5	27	9.0	20.8	0.7

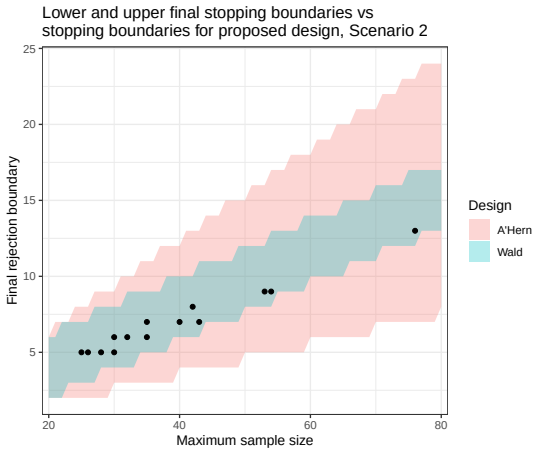
Table 2.5 Weighted loss function components $w_0ESS(p_0)$, $w_1ESS(p_1)$ and $(1-w_0-w_1)N$ for a selection of weights (w_0,w_1), for admissible design realisations for Scenario 1: $(\alpha,\beta,p_0,p_1)=(0.05,0.15,0.10,0.30)$. All values are rounded to 1 d.p.. MT: Mander and Thompson.

boundaries. However, the values of r searched over can be specified in any manner. For example, for the most complete search possible, the values of r searched over may be set to $r \in [0, N-1]$, though this would involve searching over many design realisations with unacceptable operating characteristics. Another approach to constraining r is to consider Wald's SPRT [14]. This test has no maximum sample size N. Instead, it continues until either an upper or lower boundary is reached, at which point the trial ends for a go or no go decision respectively. These boundaries may be used as a range for r. The subject of constraining r is discussed in more detail in Section 2.1.12.

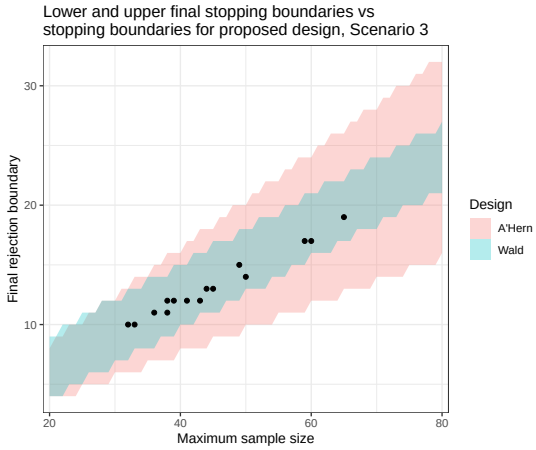
Figure 2.9 shows the final rejection boundaries for all admissible m-stage design realisations found in the design searches, alongside the range of boundaries suggested by A'Hern for single-stage designs and the lower and upper stopping boundaries for the SPRT proposed by Wald [4, 14], for each possible maximum sample size $N \in [N_{MIN}, N_{MAX}]$. As boundaries must be discrete, the lower and upper values are rounded down and up respectively. The boundaries of A'Hern and Wald have both been considered as guides for the final rejection boundary in order to reduce computation time. The range of boundaries of Wald is constant for a given set of design parameters, while that of A'Hern increases with maximum sample size N. Consequently, Wald's range is the wider range when N is low and the narrower range when N is large. In these scenarios, A'Hern and Wald's ranges are approximately equal in size at $N = 20\,(p_0 = 0.1, p_1 = 0.3$ (scenarios 1 and 2)) and at $N = 30\,(p_0 = 0.2, p_1 = 0.4$ (scenario 3)). All admissible m-stage design realisations have N great enough that Wald's range is narrower than A'Hern's. Furthermore, the final rejection boundaries of all admissible m-stage designs are within Wald's (narrow) range. As such, Wald's range is recommended as a guide for searching for final rejection boundaries as it is generally narrower than A'Hern's and therefore faster, and no admissible design realisations are likely to be missed.



(a) Scenario 1 $(\alpha = 0.05, /beta = 0.15, p_0 = 0.1, p_1 = 0.3)$



(b) Scenario 2 $(\alpha = 0.05, /beta = 0.20, p_0 = 0.1, p_1 = 0.3)$



(c) Scenario 3 $(\alpha = 0.05, /beta = 0.20, p_0 = 0.2, p_1 = 0.4)$

Fig. 2.9 Possible stopping boundaries for single-stage designs by A'Hern; lower and upper stopping boundaries for the SPRT by Wald; final rejection boundaries for admissible m-stage design realisations.

2.2.4 Scenario 2: design parameters $(\alpha, \beta, p_0, p_1) = (0.05, 0.2, 0.1, 0.3)$

Scenario 2 decreases the required power by 0.05 to $1-\beta$ =0.80 compared to scenario 1. Table 2.6 shows the optimal design realisation for each design approach across the four specified optimality criteria. The two proposed designs outperform the existing designs under p_0- and p_1-optimality. Under the p_0- and p_1-minimax criteria, the optimal Mander and Thompson designs have a lower maximum sample size ($N = 24$ vs $N = 25$ for all others), though $ESS(p_0)$ and $ESS(p_1)$ are lower for the proposed designs. Under all four optimality criteria, ESS(p_0) and ESS(p_1) of the proposed designs are lower than those of the existing designs, and the thresholds satisfy $\theta_F < 0.22$ and $\theta_E > 0.97$. Again, the ESSs of Wald's SPRT are comparable to those of the m-stage design, and again the SC design happens to be the m-stage design with the addition of an explicit interim analysis.

The design approach to which the omni-admissible design realisation belongs for each combination of weights is shown in Figure 2.10 (top left). The omni-admissible design realisation is an SC design for most combinations of weights, with exceptions where the weight of N is either close to one (as the Mander and Thompson design has the lowest N of any design) or close to zero (where m-stage is superior). The remaining plots in Figure 2.10 show the differences in loss scores between the Mander and Thompson, SC and m-stage admissible design realisations. Figures 2.10 (top right, bottom left) show that even in the region where the Mander and Thompson design is superior, the loss scores of the proposed designs are similar and as such, should be considered comparable in terms of optimality. The maximum difference in loss score in favour of the Mander and Thompson designs compared to both the SC and m-stage designs is 1.0. In Figure 2.10 (bottom right), the difference in expected loss between the admissible SC and m-stage design realisations is less than 0.9 at all points, while the range of loss scores across all admissible designs in this scenario is (11.7, 75.4), with median 20.4 (IQR [17.3, 23.2]).

	r_1	e_1	n_1	r	N	$ESS(p_0)$	$\%S_{p_0}$	$ESS(p_1)$	$\%S_{p_1}$	θ_F	θ_E
p_0-optimal											
Simon	1	–	10	5	29	15.0	1.00	26.2	1.00	–	–
MT	1	4	10	5	29	15.0	1.00	23.3	0.89	–	–
NSC	1	–	10	5	29	14.1	0.94	17.1	0.66	0.000	1.000
SC	5	–	33	7	43	11.7	0.78	13.3	0.51	0.216	0.994
m-stage	–	–	–	9	53	11.7	0.78	12.9	0.49	0.216	0.991
p_1-optimal											
Simon	2	–	18	5	25	19.9	1.00	24.6	1.00	–	–
MT	0	3	13	5	24	20.8	1.05	17.5	0.71	–	–
NSC	1	–	10	5	29	14.1	0.71	17.1	0.70	0.000	1.000
SC	3	–	19	8	43	12.5	0.63	13.0	0.53	0.163	0.980
m-stage	–	–	–	13	76	11.7	0.59	12.8	0.52	0.219	0.992
Wald	–	–	–	–	–	11.5	0.58	12.4	0.50	–	–
p_0-minimax											
Simon	1	–	15	5	25	19.5	1.00	24.6	1.00	–	–
MT	2	4	19	5	24	20.3	1.04	20.2	0.82	–	–
NSC	1	–	15	5	25	18.4	0.94	18.4	0.75	0.000	1.000
SC	0	–	9	5	25	15.3	0.79	14.6	0.59	0.058	0.973
m-stage	–	–	–	5	25	15.5	0.79	14.6	0.59	0.090	0.972
p_1-minimax											
Simon	2	–	18	5	25	19.9	1.00	24.6	1.00	–	–
MT	0	3	13	5	24	20.8	1.05	17.5	0.71	–	–
NSC	2	–	18	5	25	18.8	0.95	18.4	0.75	0.000	1.000
SC	0	–	9	5	25	15.3	0.77	14.6	0.59	0.058	0.973
m-stage	–	–	–	5	25	15.5	0.78	14.6	0.60	0.090	0.972

Table 2.6 Optimal design realisations for each design type, Scenario 2: $(\alpha, \beta, p_0, p_1) = (0.05, 0.20, 0.10, 0.30)$. For all designs, the requisite type-I error-rate and power is achieved. Columns $\%S_{p_0}$ and $\%S_{p_1}$ show ESS as a proportion of Simon's design under $p = p_0$ and $p = p_1$ respectively.

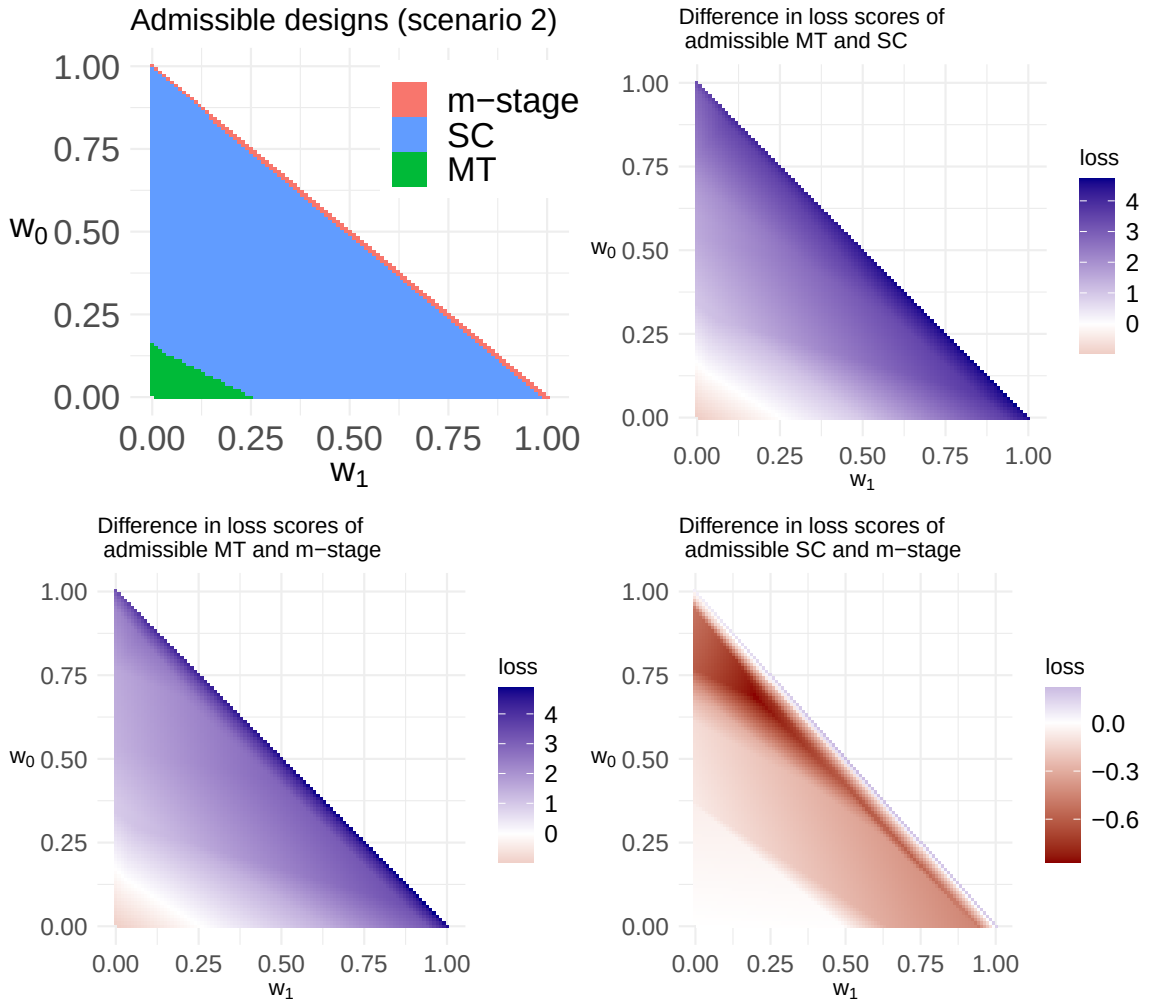


Fig. 2.10 Type of design to which the omni-admissible design realisation belongs and difference in loss scores between the following pairs of admissible design realisations: Mander and Thompson and SC (positive favours SC), Mander and Thompson and m-stage (positive favours m-stage) and SC and m-stage (positive favours m-stage). Scenario 2 $(\alpha, \beta, p_0, p_1) = (0.05, 0.20, 0.10, 0.30)$.

2.2.5 Scenario 3: design parameters $(\alpha, \beta, p_0, p_1) = (0.05, 0.2, 0.2, 0.4)$

Scenario 3 increases both p_0 and p_1 by 0.1 compared to scenarios 1 and 2, resulting in the design parameters $(\alpha, \beta, p_0, p_1) = (0.05, 0.2, 0.2, 0.4)$. Table 2.7 shows the optimal design realisations for the four optimality criteria. The proposed designs outperform the existing designs under all four criteria. In particular, the maximum sample sizes of the $p_{0/1}$-minimax proposed design realisations are lower than those of Simon's design and the NSC design. The optimal m-stage designs under p_0- and p_1-optimal have comparable ESSs to that of Wald, under both $p = p_0$ and $p = p_1$. For each design type, the admissible design realisations for p_0- and p_1-minimax are identical. The CP thresholds of the proposed designs satisfy $\theta_E > 0.98$ and $\theta_F < 0.23$ for all designs.

The design approach to which the omni-admissible design realisation belongs is shown in Figure 2.11 (left). The figure shows that each omni-admissible design realisation across the range of possible weights again belongs to either the SC or m-stage designs. The difference in expected loss between the SC and m-stage designs is shown in Figure 2.11 (right). The difference is less than 1.1 at all points, compared to the range of loss scores across all admissible designs in this scenario (15.0, 64.5), with median 26.4 (IQR [22.7, 30.4]). More context regarding relative loss scores is given in Section 2.2.3.2.

2.2.6 Effect of reduced monitoring frequency

If continuous monitoring is expected to be impractical, due to high recruitment rate, long endpoint length or for some other reason, a design permitting (stochastic) curtailment only after every block of B participants may be considered (Section 2.1.18). Such an approach can still produce savings in ESS.

Figure 2.12 shows the median, 10% and 90% quantiles for sample size as the true response rate p varies, for three design realisations. The solid lines show the median sample size, while the wide lighter ribbons show the interval of the 10th to the 90th percentile. The

	r_1	e_1	n_1	r	N	$ESS(p_0)$	$\%S_{p_0}$	$ESS(p_1)$	$\%S_{p_1}$	θ_F	θ_E
p_0-optimal											
Simon	3	–	13	12	43	20.6	1.00	37.9	1.00	–	–
MT	3	7	13	12	43	20.5	1.00	35.0	0.92	–	–
NSC	3	–	13	12	43	18.8	0.91	27.8	0.73	0.000	1.000
SC	9	–	35	13	47	15.1	0.73	20.5	0.54	0.222	0.996
m-stage	–	–	–	17	60	15.0	0.73	18.9	0.50	0.219	0.993
p_1-optimal											
Simon	4	–	18	10	33	22.3	1.00	31.6	1.00	–	–
MT	3	6	16	11	35	23.1	1.04	24.8	0.78	–	–
NSC	4	–	18	10	33	20.4	0.92	25.1	0.80	0.000	1.000
SC	13	–	44	14	47	15.8	0.71	19.1	0.60	0.146	0.986
m-stage	–	–	–	19	65	15.1	0.68	18.7	0.59	0.209	0.990
Wald	–	–	–	–	–	14.7	0.66	18.2	0.58	–	–
$p_{0/1}$-minimax											
Simon	4	–	18	10	33	22.3	1.00	31.6	1.00	–	–
MT	2	6	15	10	32	24.9	1.12	24.9	0.79	–	–
NSC	4	–	18	10	33	20.4	0.92	25.1	0.80	0.000	1.000
SC	0	–	11	10	32	21.3	0.96	20.9	0.66	0.050	0.985
m-stage	–	–	–	10	32	21.5	0.96	20.9	0.66	0.050	0.985

Table 2.7 Optimal design realisations for each design type, Scenario 3: $(\alpha, \beta, p_0, p_1) = (0.05, 0.20, 0.20, 0.40)$. For all designs, the requisite type-I error-rate and power is achieved. Columns $\%S_{p_0}$ and $\%S_{p_1}$ show ESS as a proportion of Simon's design under $p = p_0$ and $p = p_1$ respectively. p_0- and p_1-minimax designs are identical for this set of design parameters, and have been combined. MT: Mander and Thompson.

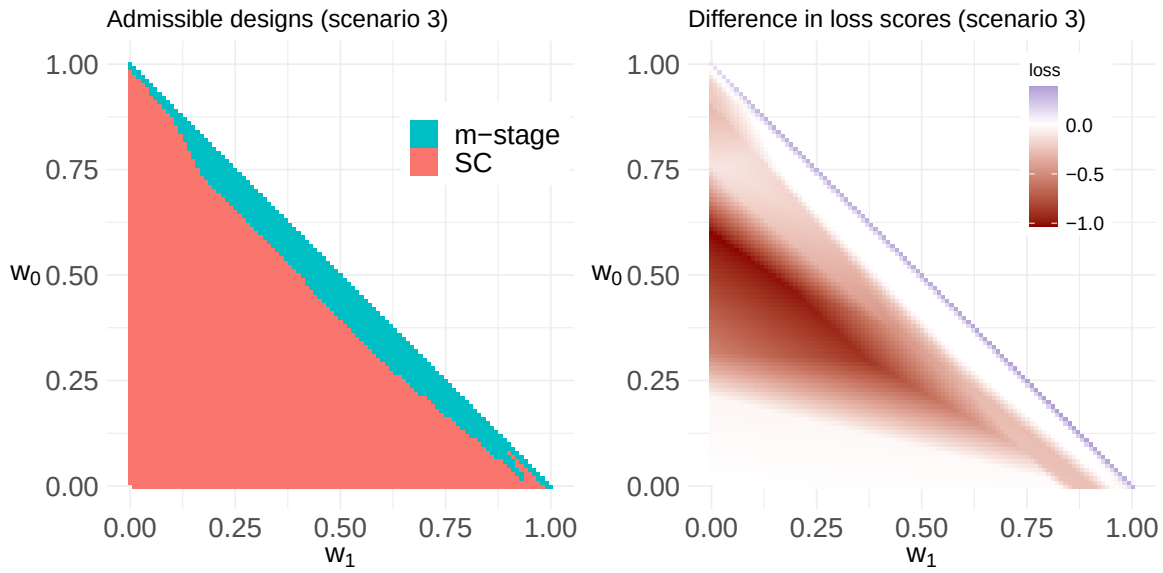


Fig. 2.11 Type of design to which the omni-admissible design realisation belongs and difference in loss scores between SC and m-stage designs (positive favours m-stage), scenario 3 $(\alpha, \beta, p_0, p_1) = (0.05, 0.20, 0.20, 0.40)$.

design realisations examined are again Simon's design used in Sharma et al. [79], with an interim analysis at $n_1 = 19$ and maximum sample size $N = 54$, and two m-stage designs that satisfy the same type-I error-rate and power requirements under the same design parameters, that is $(\alpha = 0.05, /beta = 0.1, p_0 = 0.2, p_1 = 0.4)$. One m-stage design uses continuous monitoring and was compared to Simon's design in Table 2.3 as "m-stage$_1$". The median sample size is lower than Simon's design at most points, and considerably so for $p > 0.25$. The second m-stage design realisation examined uses considerably less frequent monitoring, with an interim analysis made after every 16 participants only. With maximum sample size $N = 48$, this design realisation requires a maximum of three analyses. However, the median sample size remains lower than that of Simon's design at most points, $p \leq 0.10$, $p \geq 0.26$.

As a further example, Table 2.8 shows optimal design realisations using blocks of size four and eight, that is, permitting SC after every four or eight participants respectively, for the first scenario $(\alpha, \beta, p_0, p_1) = (0.05, 0.15, 0.10, 0.30)$. These are shown alongside the optimal design realisations for Simon's design and the m-stage design. The m-stage design may be

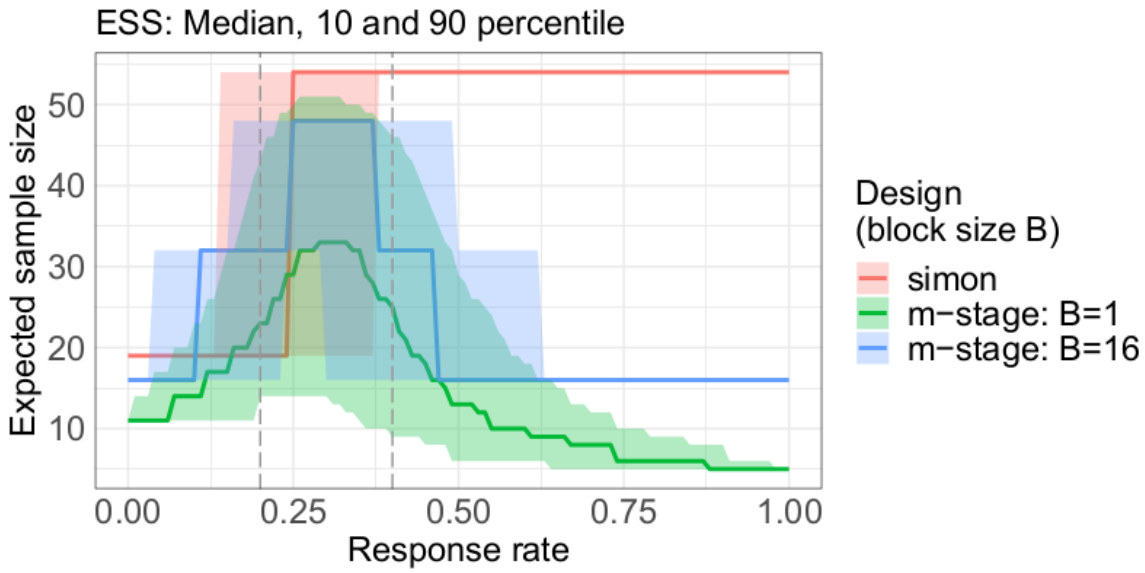


Fig. 2.12 ESS(p) for design realisations satisfying $(\alpha = 0.05, /beta = 0.1, p_0 = 0.2, p_1 = 0.4)$. Design realisations: Simon: $r_1/n_1 = 4/19, r/N = 15/54$; m-stage $(B = 1)$: $r/N = 15/52, \theta_F = 0.135, \theta_E = 0.996$; m-stage $(B = 16)$: $r/N = 14/48, \theta_F = 0.396, \theta_E = 0.991$.

considered to be equivalent to using blocks of size one. Under the p_0- and p_1-optimality criteria, the design realisations with block sizes four and eight produce considerable savings in ESS compared to using Simon's design. Under the p_0- and p_1-minimax criteria, which are combined in the table as the optimal design realisations are identical in this instance, savings in ESS are again made, with the single exception of $ESS(p_0)$ when using block size eight.

2.2.7 Estimation (scenario 1, selected)

Bias and RMSE in the response rate estimates are shown in Figures 2.13 and 2.14 for p_0-optimal design realisations for scenario 1 $(\alpha, \beta, p_0, p_1) = (0.05, 0.15, 0.10, 0.30)$, with the maximum absolute bias and RMSE shown in Table 2.9. In Simon's design and the Mander and Thompson design, the bias is close to zero for all estimators (Figure 2.13, left). For designs that employ curtailment, the bias adjusted, bias subtracted MUE and UMVUE estimators have a bias consistently close to zero, while the naïve estimator gives more biased estimates (Figure 2.13, bottom left, Figure 2.14, left). Overall, bias and RMSE is only slightly

	r_1	n_1	r	n	$ESS(p_0)$	$\%\mathrm{S}_{p_0}$	$ESS(p_1)$	$\%\mathrm{S}_{p_1}$	θ_F	θ_E
p_0-optimal										
Simon	1	11	6	35	18.3	1.00	32.3	1.00	–	–
m-stage	–	–	13	80	14.1	0.77	14.4	0.45	0.226	0.997
Block size 4	–	–	10	56	14.5	0.79	16.3	0.50	0.534	0.988
Block size 8	–	–	12	72	16.1	0.88	19.4	0.60	0.691	0.991
p_1-optimal										
Simon	2	18	5	27	20.4	1.00	26.5	1.00	–	–
m-stage	–	–	12	66	14.3	0.70	14.4	0.54	0.189	0.990
Block size 4	–	–	11	64	14.7	0.72	16.1	0.61	0.550	0.991
Block size 8	–	–	16	80	16.8	0.82	18.2	0.69	0.559	0.974
$\mathbf{p}_{0/1}$-minimax										
Simon	2	18	5	27	20.4	1.00	26.5	1.00	–	–
m-stage	–	–	5	27	18.7	0.92	16.6	0.63	0.084	0.990
Block size 4	–	–	6	32	18.8	0.92	18.7	0.71	0.194	0.984
Block size 8	–	–	6	32	21.3	1.04	21.7	0.82	0.340	0.988

Table 2.8 Selection of optimal design realisations, including stochastically curtailed designs with stopping permitted after every four and eight participants, for scenario 1: $(\alpha, \beta, p_0, p_1) = (0.05, 0.15, 0.10, 0.30)$. For all design realisations, requisite type-I error-rate and power is reached. Columns $\%\mathrm{S}_{p_0}$ and $\%\mathrm{S}_{p_1}$ show ESS as a proportion of Simon's design under $p = p_0$ and $p = p_1$ respectively.

poorer among the proposed designs than the existing designs when $p < p_1$. For greater p, the poorer estimates among the proposed designs are a result of the trial being curtailed with fewer participants compared to the existing designs. The maximum absolute bias is similar across designs, with the exception of somewhat greater bias among the proposed designs under the naïve estimator (Table 2.9).

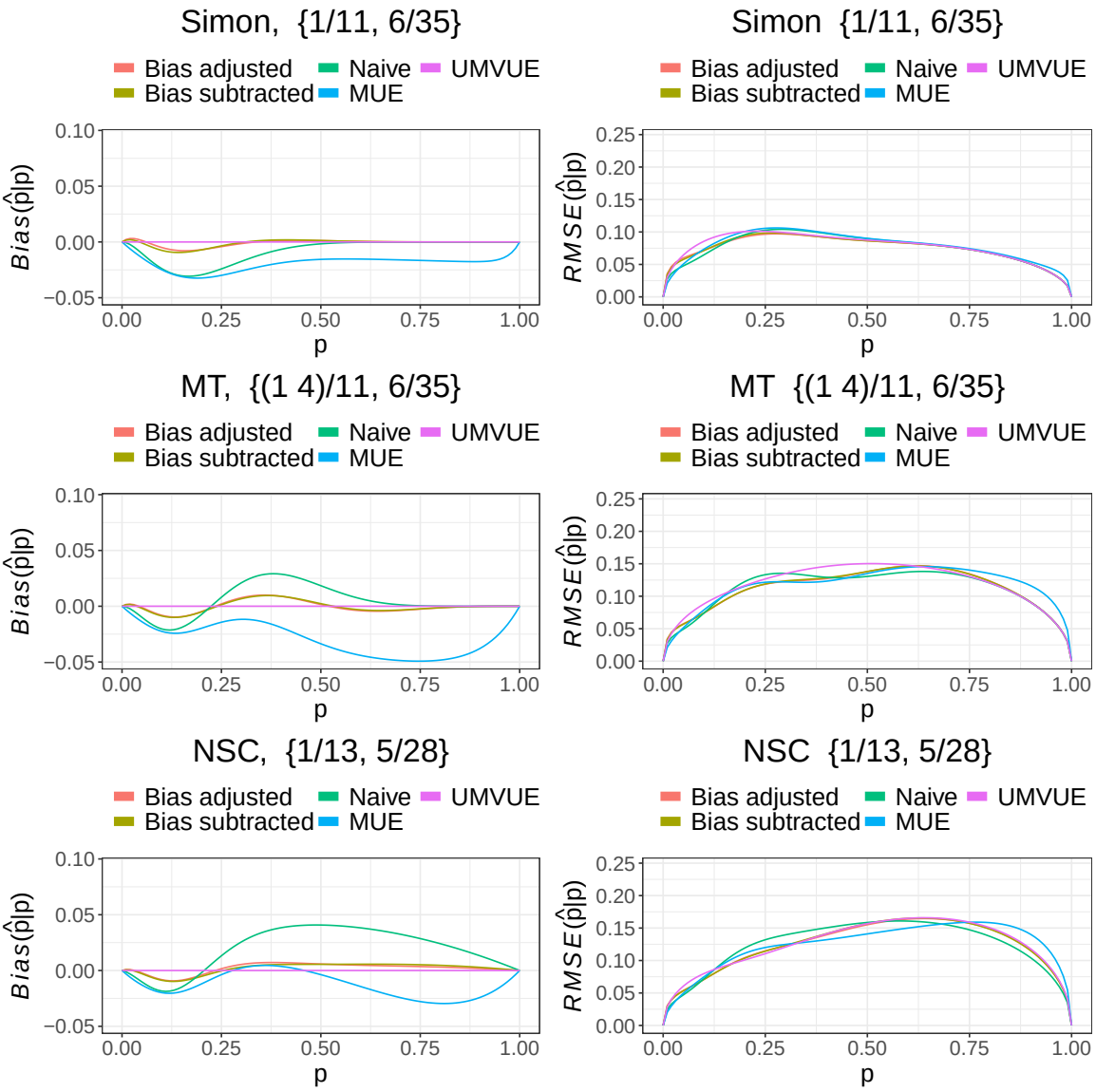


Fig. 2.13 Bias, RMSE for p_0-optimal designs, scenario 1 $(\alpha, \beta, p_0, p_1) = (0.05, 0.15, 0.10, 0.30)$.

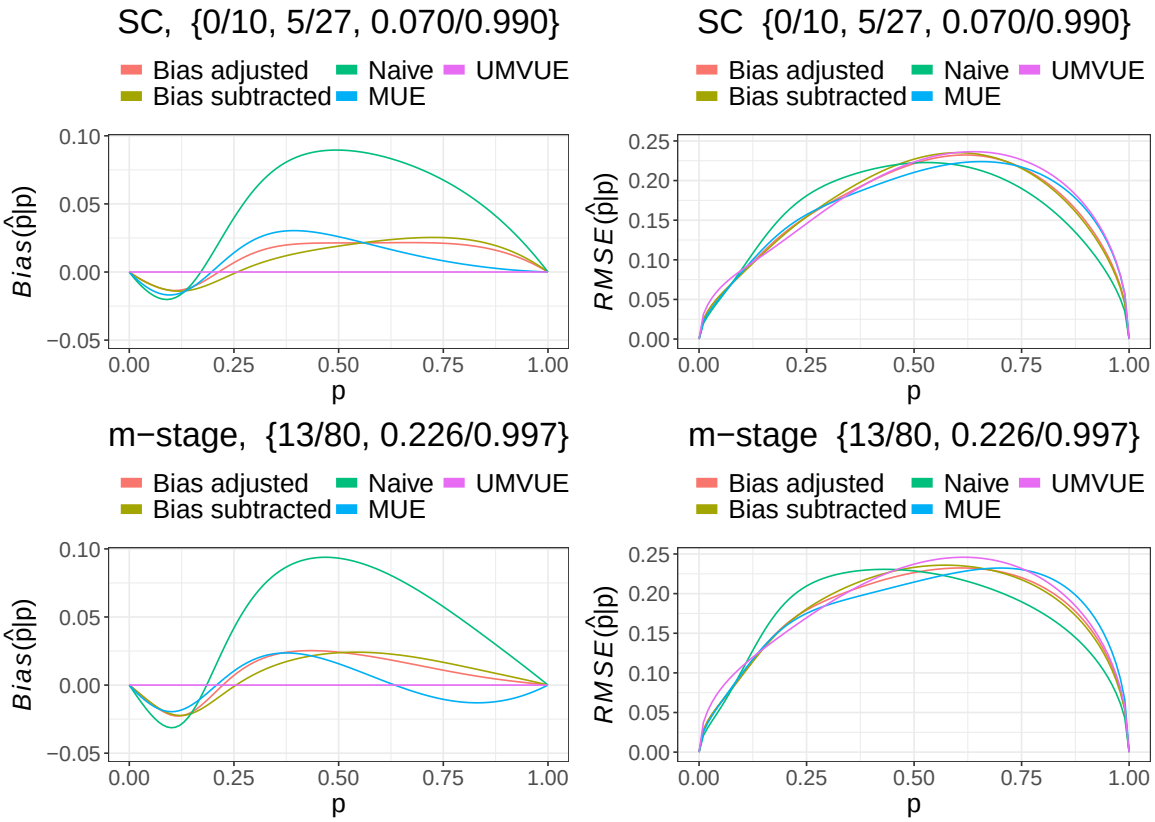


Fig. 2.14 Bias, RMSE for p_0-optimal designs, scenario 1 $(\alpha, \beta, p_0, p_1) = (0.05, 0.15, 0.10, 0.30)$, continued.

	Bias (absolute)					RMSE				
	Bias adj.	Bias subt.	Naïve	MUE	UMVUE	Bias adj.	Bias subt.	Naïve	MUE	UMVUE
Simon	0.008	0.009	0.031	0.032	4.44×10^{-16}	0.097	0.098	0.104	0.106	0.101
MT	0.010	0.010	0.029	0.049	2.22×10^{-16}	0.147	0.147	0.138	0.146	0.150
NSC	0.009	0.010	0.041	0.030	3.33×10^{-16}	0.165	0.165	0.161	0.159	0.166
SC	0.022	0.025	0.090	0.030	3.05×10^{-16}	0.232	0.235	0.223	0.224	0.236
m-stage	0.025	0.024	0.094	0.024	3.33×10^{-16}	0.232	0.236	0.231	0.232	0.246

Table 2.9 Maximum absolute bias and RMSE for various point estimators of p_0-optimal designs, scenario 1: $(\alpha = 0.05, \beta = 0.15, p_0 = 0.1, p = 0.3)$. MT: Mander and Thompson.

The RMSE of the estimates gradually increases as the degree of permitted curtailment increases, with Simon's design having the lowest RMSE and the proposed designs the greatest (Table 2.9). RMSE decreases sharply to zero as the response rate approaches one.

2.3 Discussion

In this chapter, we have introduced two new designs for binary outcome, single-arm phase II clinical trials, one based on Simon's design and one based on a single-stage design. These designs propose allowing early stopping to make a go or no go decision before the final decision would otherwise be certain.

As part of these proposed designs, this work also introduces five approaches to improving a search for an optimal or admissible design realisation that uses SC: firstly, the exact distribution of the trial outcomes is obtained, allowing the trial's operating characteristics to be known without simulation error. Secondly, a new approach is proposed for finding relevant CP thresholds when using SC, based on the CP at each point in each possible set $\{r,N\}$ or $\{r_1,n_1,r,N\}$, allowing more potential design realisations to be evaluated. Thirdly, the CP at each point in each potential design realisation is calculated taking the possibility of SC into account; it is not calculated based on an approximation that does not account for stopping due to SC. Furthermore, in the design search, type-I error-rate and power are only calculated after taking curtailment into account; no designs are discarded in advance for not achieving the required type-I error-rate and power in their uncurtailed form. Finally, the design search is undertaken using wide ranges for maximum sample size and final rejection boundary, rather than being restricted to, say, a single realisation of Simon's design. While this more expansive search could lead to extreme computation times if done in a naïve way, we present sensible heuristic constraints to reduce computational intensity. Between them, these five concepts serve dual purposes: to allow more potential designs to be examined without excessive computational intensity, and to increase the accuracy of the reported operating characteristics

of such designs. This should result in investigators being able to make a more efficient choice of design for any potential study.

The proposed designs were compared to a number of existing designs. They were compared in a real data example, where they were shown to be able to reduce the trial sample size, from 19 to 8 in one instance, and also across three scenarios with regard to the following optimality criteria: minimising ESS under $p = p_0$ or $p = p_1$ (p_0-optimal, p_1-optimal) and minimising ESS under $p = p_0$ or $p = p_1$ among designs that minimise N (p_0-minimax and p_1-minimax). With the exception of the p_0/p_1-minimax criteria in one scenario, where the proposed designs had a maximum sample size of 25 compared to 24 in an existing design, the proposed designs were superior across all criteria and scenarios. For the proposed designs, the ESSs under the p_0-optimality and p_1-optimality criteria were comparable to those obtained using Wald's SPRT [14], generally with a difference of less than a single participant in favour of Wald's SPRT. However, while Wald's SPRT may result in favourable ESSs, a design with no maximum sample size would be impractical for clinical trials, where a maximum sample size is necessary due to limited resources, population size and so on.

The proposed designs were also compared to existing designs across a combination of multiple criteria, using a weighted loss function. Employing Mander et al.'s expected loss function [37], admissible design realisations were obtained for each design approach over a grid of combinations of weights for the criteria of ESS under $p = p_0$, ESS under $p = p_1$ and maximum sample size. For each possible combination of weights, the design realisation that had the lowest expected loss across all admissible design realisations was recorded, and the type of design to which it belonged was recorded. This design realisation has been termed the omni-admissible design realisation.

Plotting the design type to which each omni-admissible design realisation belongs, for each possible combination of weights, it is shown that the proposed designs are almost always better in terms of expected loss. While the two-stage SC design can be superior to

the m-stage design, the difference is generally slight. However, we accept that increasing the maximum N searched over is likely to find design realisations with lower ESS, and the disparity between the SC and m-stage design searches in terms of maximum N searched over may be the reason why the omni-admissible design is an m-stage rather than an SC design for some combinations of weights.

We recommend that investigators focus on the m-stage design due to the decreased run time for finding admissible designs using this approach: searching for m-stage admissible designs is approximately two orders of magnitude faster than for the SC design, with a full search able to be conducted in under 60 minutes for $N \in [20, 80]$.

The effect of reducing the frequency of monitoring, was examined. It was shown that considerable savings in ESS can still be made even when employing designs with less frequent monitoring.

There may be some apprehension regarding ending a trial before the final decision is certain compared to a different design. However, the trials are powered taking this into account, in the same way that Simon's design meets the required type-I error-rate and power despite allowing stopping before the final decision is certain compared to a single-stage trial. Indeed, we have shown that Simon's designs may end for a no go decision even when the probability of success for an effective treatment is as high as 0.80. There may be particular apprehension regarding stopping to make a go decision before the final trial decision is certain compared to a single-stage or Simon's design. However, across all optimal design realisations obtained from the proposed designs in the single criterion comparisons, the threshold for stopping for a go decision was $\theta_E > 0.97$, where $\theta_E = 1$ means permitting stopping for a go decision only if the final rejection boundary is reached. Furthermore, if desired, it is possible to allow the specification of bounds to the thresholds θ_F and θ_E to ranges that are acceptable to the investigator, including $\theta_E = 1$.

In summary, this work proposes two designs for phase II, single-arm, binary outcome clinical trials, argues for using a number of approaches for finding better designs, and for using exact distributions so that the designs' operating characteristics can be obtained without simulation error. These designs have been shown to be superior to existing designs, both when considering a single optimality criterion and when considering a weighted combination of multiple criteria.

Chapter 3

Randomised binary outcome phase II trials

3.1 Methods

We present a design approach that produces randomised two-arm trials with ESSs that are far smaller than those of typical randomised two-arm trials. This design permits early stopping of a trial not only when reaching or failing to reach the required difference in the number of responses is certain, but also when it is very likely. That is, like Chapter 2, it utilises SC. The frequency of monitoring is generalised, permitting anything from a single interim analysis to monitoring that is almost continuous.

The work in this chapter is based on the paper "A stochastically curtailed two-arm randomised phase II trial design for binary outcomes" by Law et al. [89].

For a two-arm trial, let the true response rate on the control and treatment arms be p_C and p_T respectively. Our null hypothesis is as follows: $H_0 : p_T \leq p_C$. Denote by P(reject $H_0|p_C, p_T$) the probability that H_0 is rejected given response rates p_C and p_T. The nature of hypothesis testing requires us to consider what difference in treatment effect between the two arms is worth further study. To this end, we let p_0 and p_1 be response rates in the control and

treatment arms respectively, such that the treatment difference $p_1 - p_0$ is a clinically relevant difference. Define P(reject $H_0|p_C, p_T$) as the probability of rejecting H_0 given response rates of p_C on the control arm and p_T on the treatment arm. Then our design will guarantee P(reject $H_0|p_0, p_0) \leq \alpha$ and P(reject $H_0|p_0, p_1) \geq 1 - \beta$ for specified error rates α and β. Let the ESS for response rates p_C and p_T on control and treatment arms respectively be $ESS(p_C, p_T)$. Let N be the maximum total sample size and let the number of participants so far on the control and treatment arms be m_C and m_T respectively. Let $X_C(m)$ and $X_T(m)$ be the number of (binary) responses on the control and treatment arms after m participants on each arm.

3.1.1 Brief review of existing two-arm designs

Jung [46] created a design that is a two-arm analogue to Simon's design [5]. The design has a maximum sample size of $N \in 2\mathbb{Z}$ participants ($N/2$ per arm). An interim analysis takes place after $n_1 \in 2\mathbb{Z}$ participants ($n_1/2$ per arm). At this point the trial may stop for a no go decision based on the test statistic $X_T(n_1/2) - X_C(n_1/2)$, which is compared to an interim stopping boundary a_1. If $X_T(n_1/2) - X_C(n_1/2) < a_1$, the trial stops for a no go decision, otherwise it continues until the maximum N participants have been recruited. At this point, the null hypothesis is rejected if $X_T(N/2) - X_C(N/2) \geq a$, for some final rejection boundary a, otherwise it is not rejected. Type-I error-rate, power and ESS(p_0, p_0) can be calculated exactly, without requiring simulation. Note that here and below, the number of participant results in an interim analysis is given by n_1 rather than m as above. This is to distinguish the explicit, Simon-type interim analysis in a trial from the more general interim analyses that come from sequential monitoring.

Carsten and Chen [17] proposed a design that is based on the two-stage, single-arm design of Chi and Chen [9], which uses NSC and is described in Chapter 2. Similarly to Jung's design directly above, Carsten and Chen's design has a maximum sample size

of $N \in 2\mathbb{Z}$ participants ($N/2$ per arm) and an interim analysis takes place after $n_1 \in 2\mathbb{Z}$ participants ($n_1/2$ per arm). Participant results are observed in pairs, with one participant on the treatment arm and one on the control arm. The participants in each pair are matched based on some defining characteristics [17]. Each pair of results is taken together, with a "success" defined as a pair of results $\{X_{T_i}, X_{C_i}\}$ such that $X_{T_i} - X_{C_i} = 1$, that is, a response is observed on the treatment arm and a no response is observed on the control arm for some pair $i, i \in \{1, 2, \ldots, N/2\}$. The number of successes $\sum_i \mathbb{I}[X_{T_i} - X_{C_i} = 1]$, or equivalently, the number of non-successes, is the test statistic upon which each decision to reject the null hypothesis is based. By defining success in this way, the authors make no distinction between pairs of results where a response is observed on both arms, no response is observed on both arms and a response is observed on the control arm and not the treatment arm. Results for each pair are observed consecutively. After every pair in the first stage, the number of successes so far is noted. If it becomes impossible to reach some interim number of successes a_1 by $n_1/2$ pairs of results, that is, if the number of non-successes reaches $n_1/2 - a_1 + 1$, the trial ends for a no go decision. If the number of success reaches a_1 by the end of stage 1, that is, $\sum_i \mathbb{I}[X_{T_i} - X_{C_i} = 1] \geq a_1$, the trial proceeds to the second stage. Similarly, if it becomes impossible to reach some final number of successes r_2 by the end of the trial, that is, if the number of non-successes reaches $N/2 - r_2 + 1$, the trial ends and a no go decision is made. If the number of successes reaches r_2, that is, $\sum_i \mathbb{I}[X_{T_i} - X_{C_i} = 1] \geq r_2$, the trial stops for a go decision.

Chen et al. [16] also suggest a two-arm design that uses NSC. The design uses continuous monitoring, meaning that a decision regarding whether to end the trial may be taken after every participant. One consequence of this is that there is no need for any "balancing" – simple randomisation can be used in the first stage, though the stage two randomisation must be such that the final number of participants on each arm must be equal if the trial proceeds to analyse the maximum number of N participants. Success is defined as a response for a

participant on the treatment arm or a non-response for a participant on the control arm. An interim analysis is specified after n_1 participants. An interim stopping boundary for number of successes, a_1, and a final rejection boundary for number of successes, r_2, is specified. In the first stage, if it becomes impossible for the number of successes to reach a_1 by n_1 results, that is, if the number of non-successes $m_T - X_T(m_T) + X_C(m_C) \geq n_1/2 - a_1 + 1$, the trial ends for a no go decision. If it becomes certain that the number of successes will reach the interim stopping boundary after n_1 participants, that is, $X_T(m_T) + m_C - X_C(m_C) \geq n_1/2 + a_1$, the trial continues to the second stage. In the second stage, if it becomes impossible to reach the final stopping boundary r_2 by the end of the trial, that is, if $m_T - X_T(m_T) + X_C(m_C) \geq N/2 - r_2 + 1$, the trial ends and a no go decision is made. If the number of successes is certain to reach r_2, that is, if $X_T(m_T) + m_C - X_C(m_C) \geq N/2 + r_2$, the trial stops for a go decision. Chen et al. [16] also examine a single-stage version of the above design, with the interim analysis and stopping boundary omitted. Their results show that the two-stage design is superior in terms of ESS, for all comparisons made.

3.1.2 Limitations of existing designs

Jung's design [46], being a two-arm analogue of Simon's design, suffers from the same issues as Simon's design. For example, while useful for minimising ESS for inefficacious treatments, there is little saving in ESS for promising treatments. Secondly, trials continue to recruit participants even when the final go or no go decision is known with certainty.

Carsten and Chen [17] treat all three types of non-success pairs equally, that is, response on both arms, non-response on both arms and response on control arm paired with non-response on treatment arm. This is inefficient, discarding information that could otherwise contribute to the analysis. The authors also sort participants into pairs based on certain characteristics. This suggests that investigators may hope to recruit all possible participants before the trial may begin, or recruit all possible participants for the first stage before

beginning, then pausing the trial at the interim analysis to recruit all remaining participants. The authors also match the pairs of participants based on some defining characteristics. They admit that "it can be difficult" to have matching participants available at the same time, and suggest giving both treatments to all participants, so that they may serve as their own control. However, such an approach is similar to a crossover design, which is an unsuitable design for conditions that are not chronic and which has distinct disadvantages such as treatment-period interaction [44].

Carsten and Chen [17] and Chen et al. [16] use NSC, where the trial will end once it is certain that a specified number of successes will or will not be reached, and the final go or no go decision is known. However, at many possible points in binary outcome trials, such a decision may not be certain but very likely (Section 2.1.8). Allowing the trial to end at such points, in other words, using SC, could considerably reduce ESS.

The designs of Jung, Carsten and Chen and Chen et al. [46, 17, 16] use a set degree of monitoring: Jung and Chen et al. allow a decision to be made after every participant, while Carsten and Chen do so after every pair of participant results. There is no framework allowing the degree of monitoring to change based on the particular needs of a trial. These three papers find design realisations that are optimal for a single criterion, either ESS(p_0, p_0) (p_0-optimal) or ESS(p_0, p_0) among realisations that minimise N (p_0-minimax designs). There is no evaluation of the design realisations that anticipates that the treatment being tested shows promise. There is no evaluation of the design realisations that considers multiple optimality criteria.

3.1.3 Proposed two-arm design

At each interim analysis, there are three possible courses of action: stop the trial to make a go decision and reject the null hypothesis; stop the trial to make a no go decision and do not reject the null hypothesis, or continue recruitment. Define success as observing a response

on the treatment arm or a non-response on the control arm, as Chen et al. [16], with the number of successes in a trial so far defined as $S(m) := X_T(m) + m - X_C(m)$. Thus we use $S(m)$ in this chapter to denote the number of successes from m participants on each arm. We use S to denote the number of successes in general, and where the fact that $S = S(m)$ is clear. The course of action taken is determined by comparing the number of successes so far to some specified lower and upper boundaries, the calculation of which is described below. For the final analysis, it is not possible to continue the trial further, therefore either a go or no go decision is made and only a single boundary is required. Let this final stopping boundary be r. A go decision, that is, a decision to reject H_0, is made at an interim analysis if the final difference in number of responses is guaranteed to be r or greater in favour of the treatment arm. For a trial with maximum sample size N, this occurs at the end of the trial if $X_T(N/2) - X_C(N/2) \geq r$ or before the end of a trial if $X_T(m) + m - X_C(m) \geq N/2 + r$. A no go decision, that is, a decision to not reject H_0, is made as soon as the final difference in responses is guaranteed to not be r or greater in favour of the treatment arm; this occurs before the end of a trial if $m - X_T(m) + X_C(m) \geq N/2 - r + 1$. These decision rules are the NSC boundaries used by Chen et al. [16], though we relax their requirement for continuous monitoring. Jung, Carsten and Chen and Chen et al. [46, 17, 16] also include an explicit interim analysis, that is, an interim analysis at which point a go/no go decision is made regardless of whether or not the final pre-specified stopping boundary may be reached. However, in the single-arm case such an approach may result in a no go decision even when there is a high probability of correctly identifying that the null hypothesis is false (Section 2.1.4). In this chapter, H_0 is rejected if the final difference in the number of responses is greater than or equal to r, rather than strictly greater than r as in Chapter 2. This is done to align with the approaches and corresponding methods in randomised phase II trials to which we are comparing our method, in particular, Jung, Carsten and Chen and Chen et al. [46, 17, 16].

For the proposed approach, a balanced allocation between treatment and control, i.e., 1:1 randomisation, is required. The design involves frequent interim analyses, at most after every pair of observed results (one each on the control and treatment arms).

Participants are allocated to arms using block randomisation. In block randomisation, a block size is chosen and randomisation takes place within blocks such that allocation is equal between the experimental treatment and control arms. This places an upper bound on the degree of allocation imbalance that may occur [44]. Using randomised blocks and undertaking an interim analysis only at the end of each block ensures balance across the two arms at each analysis. As such, set the number of participants so far on each arm, m, to be $m = B, 2B, \ldots, N/2$, where $2B$ is the number of participants in each block. Within each block, the number of responses on each arm follows the binomial distributions $X_C(B) \sim Binom(B, p_C)$ and $X_T(B) \sim Binom(B, p_T)$, for a fixed number of participants B per block per arm. Stopping is permitted after each block. Permitting stopping after each block is a sensible approach to early stopping, allowing trials to end early but without the need to make a decision after every participant or pair of participants, which may be impractical in large randomised two-arm trials. Furthermore, the block size may be chosen to suit the resources of the trial, with smaller block sizes allowing decisions to be made earlier and larger block sizes requiring fewer early stopping decisions. It may be possible to undertake continuous monitoring and update a trial design after each participant [21]. However, in some circumstances this is not possible, and the proposed design does not require such a degree of monitoring. Requiring continuous monitoring may increase the occurrence of delayed responses, described in Section 2.1.5 and suggested as future work (Section 5.4). The flexible framework permits a wide range of degrees of monitoring, from a small number of interim analyses to monitoring that is almost continuous.

The test statistic we will use to determine whether to reject the null hypothesis is the difference in the number of responses between the arms, $X_T(m) - X_C(m)$, though other valid

test statistics exist. In particular, the decision of whether to study a treatment further may depend not solely on rejection of the null hypothesis, but also on other factors. For example, a design may explicitly require a minimum effect size estimate before permitting further study [90–92]. Some such methods involve sculpting of the rejection region. However, the true response rates p_C and p_T may not be equal to the specified response rates p_0 and p_1. If so, sculpting the rejection region can lead to underestimating the type-I error-rate [93]. Sensitivity to such deviations will be examined in the Section 3.2.3 and we discuss this issue further in the Discussion (Section 5.3). Moreover, it is usually possible to design a trial by specifying the improvement in response rate that would be clinically worthwhile [44].

We present a design approach that permits early stopping of a trial not only when reaching or failing to reach the required difference in the number of responses is certain, but also when it is very likely, that is, using SC. SC has previously been applied to single-arm binary trials as detailed in Chapter 2 [13, 12]. However, SC has not previously been applied to two-arm binary outcome trials. Another clinical trial characteristic that is utilised in our approach is block randomisation.

We note that other two-arm approaches have been proposed [94–99]. However, it is impractical to compare all randomised two-arm designs, and so our approach will be compared only to Jung's design [46], as this is the two-arm analogue to the popular Simon design, and Carsten and Chen and Chen et al.'s designs [17, 16], as these designs use curtailment and as such are similar to our approach. Table 3.1 shows the main differences between the designs to be compared. As can be seen, our approach uses a test statistic that has been used in other approaches, allows both stochastic and non-stochastic curtailment, allows early stopping based on how likely trial success is, and allows a flexible number of interim analyses.

Approach	NSC	SC	Early stopping for go decision	Early stopping w/out curtailment	No. stopping decisions	Test statistic
Jung	No	No	No	Yes	2	$X_T - X_C$
Carsten and Chen	Yes	No	Yes	Yes	$N/2$	$\sum_{i=1}^{m} \mathbb{I}(X_{Ti} = 1, X_{Ci} = 0)$
Chen et al.	Yes	No	Yes	Yes	N	$X_T - X_C$
Block design	Yes	Yes	Yes	No	$N/2B$	$X_T - X_C$

Table 3.1 Characteristics of the two-arm designs to be compared. m: number of participants per arm so far; B: number of participants per arm per block.

3.1.4 Conditional power in the two-arm setting

For a two-arm design, define the conditional probability, $CP(p_C, p_T, S(m), m)$, as the probability of rejecting H_0 conditional on observing $S(m)$ successes after m participants on each arm assuming some response rates p_C and p_T, with r and N fixed. Setting $p_C = p_0, p_T = p_1$ gives the conditional power $CP(p_0, p_1, S(m), m)$. For the purposes of the proposed approach, the only conditional probability of interest is the conditional power, and so in this chapter $CP(S(m), m)$ will refer solely to conditional power $CP(p_0, p_1, S(m), m)$, while the abbreviation CP will refer to conditional power in general. $CP(S(m), m)$ is calculated using p_0 and p_1, that is, there is no re-estimation of response rates. SC in this design is based on CP, though we acknowledge that other approaches are available [10].

3.1.4.1 Calculating conditional power under NSC

When $m - X_T(m) + X_C(m) \geq N/2 - r + 1$, it is no longer possible for the null hypothesis to be rejected, and so $CP(S(m), m)$ is equal to zero. Conversely, when $X_T(m) + m - X_C(m) \geq N/2 + r$, rejection of the null hypothesis is guaranteed, and so $CP(S(m), m)$ is equal to one. For a block design with no explicit interim analysis and using NSC but not SC, $CP(S(m), m)$ can be written recursively as

$$CP(S(m),m) = \left\{ \begin{array}{ll} 0, & \text{if } m - X_T(m) + X_C(m) \geq N/2 - r + 1 \\ D, & \text{if } m - X_T(m) + X_C(m) < N/2 - r + 1 \\ & \quad \text{and } X_T(m) + m - X_C(m) < N/2 + r \\ 1, & \text{if } X_T(m) + m - X_C(m) \geq N/2 + r \end{array} \right\}, \tag{3.1}$$

where

$$D = \sum_{i=0}^{2B} P(i,B|p_0,p_1) CP(S+i,m+B)$$

and $P(i,B|p_C,p_T) = P(S(B) = i|p_C,p_T) = P\left(X_T(B) + B - X_C(B) = i|p_C,p_T\right)$, the probability of observing i successes from a block containing B participants on each arm, given response rates p_C and p_T. CP$(S(N),N)$ = 1 for $S(N) \geq r$, 0 otherwise.

3.1.4.2 Calculating conditional power under SC

SC entails ending a trial not only at any point where CP is equal to zero or one, but also for a no go decision at any point where $0 < CP(S(m),m) < \theta_F$ or for a go decision at any point where $\theta_E < CP(S(m),m) < 1$, for fixed thresholds $(\theta_F, \theta_E) \in [0,1]$ such that $\theta_F < \theta_E$.

To incorporate SC, only slight changes to Equation (3.1) are required:

$$CP(S(m),m) = \left\{ \begin{array}{ll} 0, & \text{if } m - X_T(m) + X_C(m) \geq N/2 - r + 1 \text{ or } D < \theta_F \\ D, & \text{if } m - X_T(m) + X_C(m) < N/2 - r + 1 \\ & \quad \text{and } X_T(m) + m - X_C(m) < N/2 + r \\ & \quad \text{and } \theta_F \leq D \leq \theta_E \\ 1, & \text{if } X_T(m) + m - X_C(m) \geq N/2 + r \text{ or } D > \theta_E \end{array} \right\}, \tag{3.2}$$

Equations (3.1) and (3.2) are recursive as the CP at a given point, $CP(S(m),m)$ say, is dependent on the CP at "future" points $CP(S+i,m+B), i = 0,1\ldots,2B$. Under SC (Equation (3.2)), CP values lower than θ_F are set to zero and CP values greater than θ_E are set to one, as early stopping occurs at such points. These equations are analogous to the calculation of CP in single-arm trials introduced in Section 2.1.7.1 (Equations (2.3) and (2.5)). Calculating CP in this manner accounts for the possibility of early stopping due to SC. Thus it ensures that the operating characteristics are known exactly, and that any decision to continue the trial is done so knowing exactly what degree of uncertainty remains about whether to reject H_0. By calculating the CP at each point $(S(m),m), m = B,2B,\ldots,N/2, S = 0,1,\ldots,2m$, the stopping boundaries for the conclusion of each block are obtained. Knowing in advance which points will, if reached, result in early stopping means that the exact distribution of the trial's outcomes are known. Furthermore, calculating CP without error at each point, rather than using an approximation, prevents decisions being made based on a CP with unknown error.

Let any particular example of a trial created using our approach be characterised by $\{r,N,B,\theta_F,\theta_E\}$, and denote any such example to be a "realisation" of our design. Each design realisation has explicit lower and upper limits for CP, θ_F and θ_E, one of which must be reached before the trial may end. For existing design approaches that permit early stopping to reject H_0 under NSC only, such as Carsten and Chen and Chen et al. [17, 16], the equivalent lower and upper limits for stopping the trial are $\theta_F = 0$ and $\theta_E = 1$ respectively, and cannot be altered. That is, the CP must equal zero for a no go decision to be made and must equal one for a go decision to be made.

3.1.5 Design search

The paramount requirements of a design realisation are that the desired type-I error-rate α and power $1-\beta$ are satisfied, that is, $\alpha^* = \text{P(reject } H_0|p_0,p_0) \leq \alpha$ and $1-\beta^* =$

P(reject $H_0|p_0, p_1) \geq 1-\beta$. As in the single-arm case, designs that satisfy these requirements are denoted feasible. We wish to consider only feasible designs. It is worthwhile to compare the ESS of feasible design realisations. ESS for a given design is obtained by finding all possible points at which the trial will end, then multiplying the number of participants so far at those points by the probability of reaching such points. For response rates p_C on the control arm and p_T on the treatment arm, this is

$$ESS(p_C, p_T) = \sum_{j=1}^{N/2B} \sum_{i=0}^{2jB} 2jB \; P(i, jB|p_C, p_T)\mathbb{I}(CP(i, jB) \in \{0,1\}).$$

Our interest lies in ESS under $p_C = p_T = p_0$, $ESS(p_0, p_0)$ and ESS under $p_C = p_0, p_T = p_1$, $ESS(p_0, p_1)$.Again as in the single-arm case, design realisations that are superior to all others for any combination of multiple optimality criteria are described as admissible. The term "admissible" has previously been used with respect to two-arm designs [46, 37], and these design realisations are our subject of interest. It is the admissible design realisations of our proposed approach that will be compared, both to one another and to admissible design realisations of other approaches.

In order to find admissible design realisations, a search of possible designs is undertaken. The block size $2B$, desired type-I error-rate α and power $1-\beta$ are specified in advance, as is an upper limit for maximum sample size, N_{MAX}, as may a range for the final rejection boundary r. Choice of r is discussed in the single-arm case (Section 2.2.3.3). Also specified in advance are a maximum lower limit and minimum upper limit for CP, denoted $\theta_{F_{MAX}}$ and $\theta_{E_{MIN}}$, so that the design search takes place only over combinations $\{r, N, B, \theta_F, \theta_E\}$ that satisfy $\theta_F \leq \theta_{F_{MAX}}$ and $\theta_E \geq \theta_{E_{MIN}}$. For all results that follow, $\theta_{F_{MAX}}$ was set equal to p_1, that is, a trial's CP threshold for ending for a no go decision may not be greater than the anticipated response rate on treatment, or $\theta_F \leq p_1$. This is a pragmatic choice: it is a sensible constraint to not consider a no-go decision if the current conditional probability of trial success is greater than the probability of observing a response in a single participant allocated

to a treatment with response rate p_1. A fixed value of 0.7 was chosen for $\theta_{E_{MIN}}$, meaning that a trial's CP threshold for ending for a go decision may not be less than 0.7, that is, $\theta_E \geq 0.7$. This value was considered a reasonable minimum probability for making a go decision, though in practice this value may be determined in collaboration with investigators. If an investigator wishes to allow early stopping for a go decision only when CP is high, then this may be set to, for example $\theta_{E_{MIN}} = 0.95$, or even $\theta_{E_{MIN}} = 1$ if an investigator wishes to permit early stopping for a go decision only when reaching the final stopping boundary r is certain. The final value that may be specified is the maximum number of (θ_F, θ_E) combinations to be tested per unique $\{r, N\}$. This is further explained below.

Searches were undertaken for two block sizes, $2B \in \{2, 8\}$, and for five values of control arm response rate, $p_0 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, with $p_1 = p_0 + 0.2$ in each instance. This resulted in ten searches overall. These block sizes were chosen to examine to what extent the operating characteristics change when the degree of monitoring is reduced considerably. The searches had the following parameters: $N_{MAX} = 120$, $r \in \{0, 1, \ldots, \lceil Np_1 \rceil\}$, $\alpha = 0.15, \beta = 0.2$ (as in Table 1 of Jung [46]), $\theta_{F_{MAX}} = p_1, \theta_{E_{MIN}} = 0.7$ and maximum number of (θ_F, θ_E) combinations 10^6. The maximum sample size, CP limits and maximum number of (θ_F, θ_E) combinations were pragmatic choices, balancing the desire to search over as many design realisations as possible against computational intensity. Each trial, with design parameters $\{r, N, B, \theta_F, \theta_E\}$, was evaluated to obtain $\alpha^*, 1 - \beta^*, ESS(p_0, p_0)$ and $ESS(p_0, p_1)$.

3.1.5.1 Searching over CP thresholds θ_F and θ_E

For any given set $\{r, N, B\}$, each possible combination of successes, S, and participants so far, $2m$, has an associated CP. As θ_F and θ_E vary, the operating characteristics of a trial are only certain to change when θ_F or θ_E become greater than or less than one of the possible CP values in the trial. As such, we have chosen to vary θ_F and θ_E over the trial-specific CP values rather than searching over uniform distributions of θ_F and θ_E, as in the proposed single-arm

designs. That is, $(\theta_F, \theta_E) \in \{CP(S(m), m), m = B, 2B, \ldots, N/2, S = 0, \ldots, r : \theta_F < \theta_E, \theta_F \leq \theta_{F_{MAX}}, \theta_E \geq \theta_{E_{MIN}}\}$.

For large sample sizes, the number of unique CP values and consequently, the number of possible (θ_F, θ_E) combinations to be searched over may be great. However, certain aspects of our design approach can ameliorate this to some degree. Firstly, in many cases, the CP is equal to zero or one. Secondly, by only permitting stopping after every $2B$ participants, we need only consider the CP values that occur at the conclusion of each block, that is, after $B, 2B, \ldots, N/2$ participants on each arm. Finally, there are the user-defined limits set above: $\theta_F \leq \theta_{F_{MAX}}$, $\theta_E \geq \theta_{E_{MIN}}$. These three aspects reduce the number of unique CP values for each trial design. Nevertheless, the number of possible (θ_F, θ_E) combinations still increases rapidly with N (Section 2.1.11). In the single-arm case, we specified an upper bound Θ for the number of CP values $|\theta|$. Here, this has been superseded by a more direct approach and we specify a maximum number of ordered pairs (θ_F, θ_E) that may be examined per $\{r, N, B\}$. As stated above, the limit chosen was 10^6, meaning that for each $\{r, N, B\}$ combination, at most 10^6 combinations of (θ_F, θ_E) are examined. When there are more than 10^6 possible combinations, the (unique) CP values are ordered from smallest to greatest, then every other value is removed, excluding zero and one. This thinning is repeated until the number of possible combinations remaining is not greater than 10^6. As the number of CP values $|\theta|$ becomes large, the resulting number of ordered pairs is approximately equal to $|\theta|^2/2$, which allows simple comparison between maximum number of CP values Θ and maximum number of ordered pairs. Thus, this is a greater restriction than used in the single-arm case, where the maximum number of CP values was 10^6. As the distribution of CP values in a trial is not uniform (Section 2.1.8), this approach is less likely to miss potential designs than simply searching over a uniform distribution of CP values.

3.1.5.2 Design search: algorithms/pseudocode

We now describe the design search in detail, first in words and then using pseudocode. The design search is similar to the single-arm design search at a high level: a family $\mathscr{R}$ of sets containing all $\{r,N,B\}$ are found then CP values are obtained for each $\{r,N,B\}$, from which ordered pairs (θ_F,θ_E) are used to find design realisations. However, the functions used at many steps must be altered to account for two-arm data. Moreover, the procedure for finding (θ_F,θ_E) is more complex. In the single-arm case, stopping decisions are based on the number of responses $S(m)$ observed after m participants. A particular point in an uncurtailed single-arm trial $(S(m),m)$ may be reached via many different paths. However, all paths are equally likely, having probability $p^{S(m)}(1-p)^{m-S(m)}$, with total probability $\binom{m}{S(m)}p^{S(m)}(1-p)^{m-S(m)}$ for some response rate p. In contrast, in the two-arm case, stopping decisions are based on the number of successes after m participants per arm, that is, $S(m) = X_T(m/2)+m/2-X_C(m/2)$. A given number of successes can be reached via paths of differing probabilities. For example, two successes after four participants may be due to observing any of three different possibilities: two responses on the treatment arm and two responses on the control arm $(X_T(2)=2,X_C(2)=2)$, with probability $p_T^2p_C^2$; one response on the treatment arm and one response on the control arm $(X_T(2)=1,X_C(2)=1)$, with probability $p_T(1-p_T)p_c(1-p_C)$; zero responses on the treatment arm and zero responses on the control arm $(X_T(2)=0,X_C(2)=0)$, with probability $(1-p_T)^2(1-p_C)^2$. Consequently, finding the operating characteristics of two-arm trials is slower than in single-arm trials. This, combined with the larger sample size required in two-arm trials compared to single-arm trials, means that reducing computational intensity is important. In addition to restricting the range of r for each N searched over and placing constraints on both the total number of ordered pairs (θ_F,θ_E) and the values themselves $(\theta_F \leq \theta_{F_{MAX}}, \theta_E \geq \theta_{E_{MIN}})$, we may also restrict how the design search explores the possible (θ_F,θ_E) ordered pairs. In particular, we provide two options for exploring this space, one of which examines all ordered pairs

that produce feasible design realisations at the expense of speed, while the other examines fewer ordered pairs but is quicker. Given the similarity of the design search to the single-arm case otherwise, it is only necessary to provide details of this aspect of the search, shown in Algorithms 2 and 3.

Algorithm 2: Slow search over CP values θ for a single $\{r, N, B\} \in \mathscr{R}$

```
k ← 1;
ordered.pairs.matrix ← call findOrderedPairs(θ);
thetaE.vals ← call unique(ordered.pairs.matrix[,2]);
thetaE.vals ← call sortDecreasing(thetaE.vals);
no.thetaE.vals ← call length(thetaE.vals);
for i = 1 to no.thetaE.vals do                    // for each unique θ_E, obtain a
 corresponding vector of θ_F values from the ordered pairs matrix:
    j ← 1;
    for row = 1 to nrow(ordered.pairs.matrix) do
        if ordered.pairs.matrix[row, 2] = thetaE.vals[i] then
            current.thetaF.vec[j] ← ordered.pairs.matrix[row, 1];
            j ← j+1;
        end
    end
    for q = 1 to length(current.thetaF.vec) do
        design.OCs.matrix[k,] ← call findOCs(r, N, B, θ_F=current.thetaF.vec[q],
          θ_E=thetaE.vals[i], ...);
        pwr ← call findPower(design.OCs.matrix[k,]);
        k ← k+1;
        if pwr < power then
            break
        end
    end
end
output ← call discardDominatedDesigns(design.OCs.matrix)
```

For increased clarity, the faster method for exploring the ordered pairs can first be described in words as follows:

- Create two vectors $\boldsymbol{\theta_F}$ and $\boldsymbol{\theta_E}$ from the CP values θ, satisfying $\boldsymbol{\theta_F} : \theta \leq \theta_{F_{MAX}}$ and $\boldsymbol{\theta_E} : \theta \geq \theta_{E_{MIN}}$.

- Bisect $\boldsymbol{\theta_F}$ and $\boldsymbol{\theta_E}$ to find the central value of each, and find the design operating characteristics using these values.

- If this design realisation is not feasible, bisect the vectors again, between the minimum and current values of $\boldsymbol{\theta_F}$, that is, decreasing θ_F, and between the current and maximum values of $\boldsymbol{\theta_E}$, that is, increasing θ_E, and find the design operating characteristics for using these values. Continue bisecting until either $(\theta_F = 0, \theta_E = 1)$ or a feasible design is found.

- If no feasible design is found by $(\theta_F = 0, \theta_E = 1)$, end and move on to the next set $\{r, N, B\}$.

- If a feasible design is found for some (θ_F, θ_E):

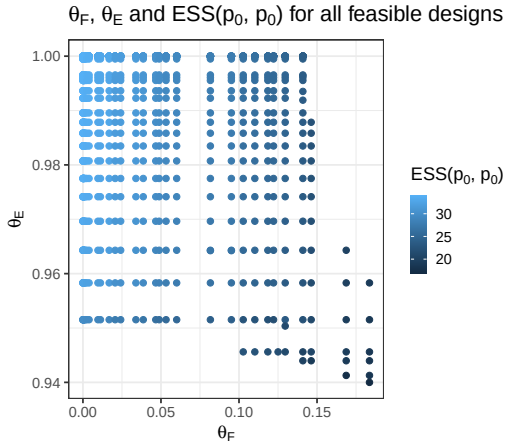
  - Define $\theta_{E_{MAX}}$ as the current value of θ_E.
  - Find design operating characteristics for $(\theta_F, \theta_{E_{MAX}})$, using sequentially increasing values in $\theta_F \in \boldsymbol{\theta_F}$, stopping when either a non-feasible design is reached or $\theta_F = \theta_{F_{MAX}}$.
  - Upon stopping, define $\theta_{F_{MIN}}$ as the current value of θ_F
  - Find design operating characteristics for every (θ_F, θ_E) such that $\theta_{F_{MIN}} \leq \theta_F \leq \theta_{F_{MAX}}$ and $\theta_{E_{MIN}} \leq \theta_E \leq \theta_{E_{MAX}}$, $\theta_F \in \boldsymbol{\theta_F}$, $\theta_E \in \boldsymbol{\theta_E}$.

Justification for such a procedure can be seen in Figure 3.1, which shows plots of (θ_E, θ_F), ESS(p_0, p_0) and ESS(p_0, p_1) for all feasible designs found with specified design parameters. The Figures show how both ESS(p_0, p_0) decreases as θ_F increases and conversely ESS(p_0, p_1) decreases with θ_E, for a fixed N. Furthermore, most feasible designs exist at extreme values of θ_F (low) and θ_E (high). With this in mind, avoiding searching at these extremes should

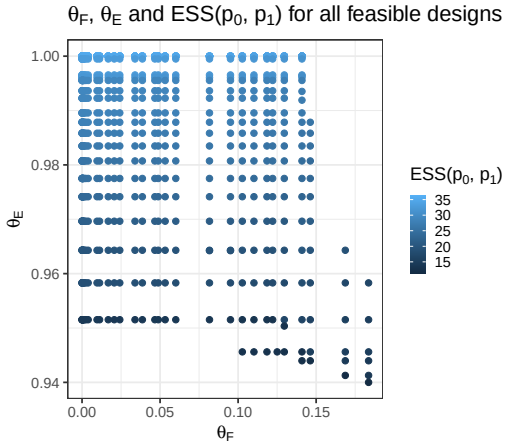
Algorithm 3: Fast search over CP values for a single $\{r, N, B\} \in \mathscr{R}$

$\boldsymbol{\theta_F} \leftarrow$ call subset(θ, max=$\theta_{F_{MAX}}$);
$\boldsymbol{\theta_E} \leftarrow$ call subset(θ, min=$\theta_{E_{MIN}}$);
$a_0 \leftarrow 1$;
$b_0 \leftarrow$ length($\boldsymbol{\theta_F}$);
$d_0 \leftarrow$ ceiling($(b_0 - a_0)/2$);
$a_1 \leftarrow 1$;
$b_1 \leftarrow$ length($\boldsymbol{\theta_E}$);
$d_1 \leftarrow$ ceiling($(b_1 - a_1)/2$);
feasible $\leftarrow$ FALSE;
$k \leftarrow 1$;
while $b_0 - a_0 > 1$ and $b_1 - a_1 > 1$ and feasible=FALSE **do**
  temp.output $\leftarrow$ call findDesignOCs($\boldsymbol{\theta_F}[d_0]$, $\boldsymbol{\theta_E}[d_1], \ldots$);
  feasible $\leftarrow$ call feasibleT1.Power(temp.output);
  if feasible=FALSE **then**
    $b_0 \leftarrow d_0$;
    $d_0 \leftarrow a_0 +$ ceiling($(b_0 - a_0)/2$);
    $a_1 \leftarrow d_1$;
    $d_1 \leftarrow a_1 +$ ceiling($(b_1 - a_1)/2$);
  else
    $\theta_{E_{MAX}} \leftarrow \boldsymbol{\theta_E}[d_1]$;
    while $d_0 <$ length($\boldsymbol{\theta_F}$) and feasible=TRUE **do**
      $d_0 \leftarrow d_0 + 1$;
      design.OCs.matrix[k,] $\leftarrow$ call findDesignOCs($\boldsymbol{\theta_F}[d_0]$, $\theta_{E_{MAX}}, \ldots$);
      feasible $\leftarrow$ call feasibleT1.Power(design.OCs.matrix[k,]);
      $k \leftarrow k + 1$;
    end
    $\theta_{F_{MIN}} \leftarrow \boldsymbol{\theta_F}[d_0 - 1]$;
  end
end
if exists($\theta_{F_{MIN}}$) **then**
  $\boldsymbol{\theta_F} \leftarrow$ call subset(θ, min=$\theta_{F_{MIN}}$, max=$\theta_{F_{MAX}}$);
  $\boldsymbol{\theta_E} \leftarrow$ call subset(θ, min=$\theta_{E_{MIN}}$, max=$\theta_{E_{MAX}}$);
  for i in 1 to length($\boldsymbol{\theta_F}$) **do**
    for j in 1 to length($\boldsymbol{\theta_E}$) **do**
      design.OCs.matrix[k,] $\leftarrow$ call findDesignOCs($\boldsymbol{\theta_F}[i]$, $\boldsymbol{\theta_E}[j], \ldots$);
      $k \leftarrow k + 1$;
    end
  end
end
output $\leftarrow$ call discardDominatedDesigns(design.OCs.matrix)

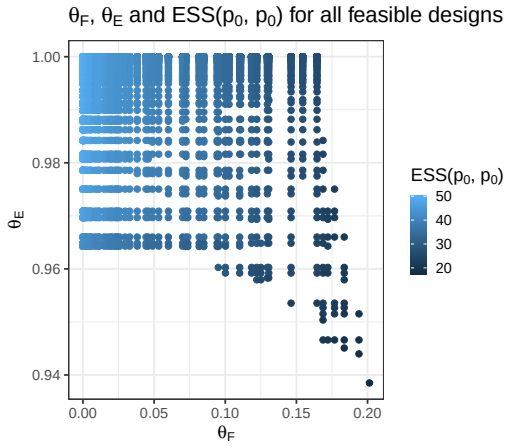
speed up the design search while still finding design realisations with favourable ESS(p_0, p_0) and ESS(p_0, p_1). In these examples, this search method is approximately one order of magnitude faster than the more simple search when allowing a maximum of 10^6 ordered pairs (6 seconds vs. 53 seconds for $\{r = 3, N = 40, B = 2\}$, 35 seconds vs. 500 seconds for $\{r = 5, N = 60, B = 2\}$). In both examples, the same best design realisation was found (as each example contained a design single realisation that minimised both ESS(p_0, p_0) and ESS(p_0, p_1)), though this is not guaranteed in general.



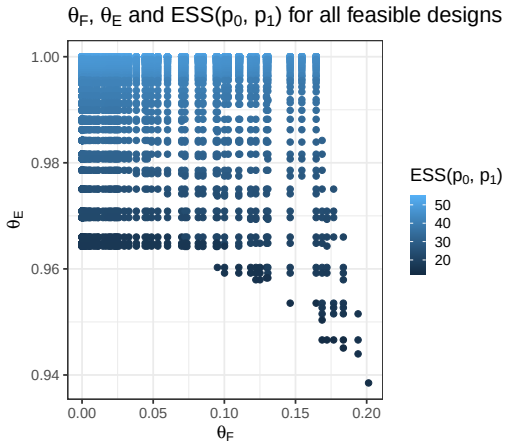
(a) ESS(p_0, p_0) for $\{r = 3, N = 40, B = 1\}$



(b) ESS(p_0, p_1) for $\{r = 3, N = 40, B = 1\}$



(c) ESS(p_0, p_0) for $\{r = 5, N = 60, B = 1\}$



(d) ESS(p_0, p_1) for $\{r = 5, N = 60, B = 1\}$

Fig. 3.1 Plots showing (θ_F, θ_E), ESS(p_0, p_0) and ESS(p_0, p_1) for all feasible designs with $\alpha = 0.15$, $\beta = 0.2$, $p_C = 0.1$, $p_T = 0.4$, $B = 1$, for two selected sets of $\{r, N, B\}$.

3.1.5.3 Design search: existing designs

Regarding searches using existing designs, admissible designs for Jung's design [46] were found using the R package `ph2rand` [74, 100]. Admissible designs for the designs of Carsten and Chen and Chen et al. [17, 16] were found using simulation as suitable code was not available: all design combinations $\{r_1, n_1, r, N\}$ such that $N \in [10, 200]$, with restrictions $r_1 \leq n_1$, $r_1 \leq r$, were obtained, where r_1 denotes the interim stopping boundary that must be reached after n_1 participants for the trial to continue. For each design, α^* and β^* were initially estimated using 100 simulated datasets. Designs with $\alpha^* > 0.25$ (that is, $\alpha + 0.1$) or $\beta^* > 0.3$ (that is, $\beta + 0.1$) were discarded. For the remaining designs, α^*, β^*, $ESS(p_0, p_0)$ and $ESS(p_0, p_1)$ were estimated using 10,000 simulated datasets. Designs with $\alpha^* > 0.15$ or $\beta^* > 0.2$ were discarded, as were dominated designs, leaving a set of admissible designs for both approaches. To avoid confusion, the design of Carsten and Chen [17] will be described in the Results section as "Carsten".

3.1.6 The loss function

The concept of using a loss score in the form of a weighted sum of optimality criteria to compare trial designs was used to compare designs in Chapter 2, and has been used previously [46, 37]. We use the approach of Mander et al. [37], extended to the two-arm case. The loss score of a two-arm design realisation is defined as

$$L = w_0 ESS(p_0, p_0) + w_1 ESS(p_0, p_1) + (1 - w_0 - w_1)N,$$

where $w_0, w_1 \in [0, 1]$ and $w_0 + w_1 \leq 1$. For all combinations of weights w_0 and w_1, the loss score is compared across admissible design realisations from different approaches. To further compare admissible design realisations produced by different approaches, we note the design realisation with the lowest loss score for each combination of weights (among all

design approaches). This design is termed the omni-admissible design, as in the single-arm case. The omni-admissible design is deemed to be the best-performing design realisation for that combination of weights. The design type of each omni-admissible design is obtained for each combination of weights. The results are plotted, to visualise what approach performs best for each weighting of optimality criteria.

3.1.7 Comparison of proposed and existing designs: summary

We compare our proposed design, using blocks of size two and size eight, to existing designs using the weighted combination of multiple optimality criteria described above. Optimal design realisations for a number of single optimality criteria are found both a series of response rates (p_0, p_1) and for a real-life example. We examine the effect of the true response rates p_C, p_T deviating from the specified values of p_0, p_1.

3.2 Results

3.2.1 Comparing design approaches using multiple criteria

All results are based on the operating characteristics $(\alpha, \beta) = (0.15, 0.20)$, as used in Table 1 of Jung [46], and the range $p_0 = \{0.1, \ldots, 0.5\}, p_1 = p_0 + 0.2$. To address the case of greater response rates, a real-life example is investigated where $p_0 = 0.70, p_1 = 0.85$ [101].

Figure 3.2 shows the design approach to which the omni-admissible design belongs, that is, the design realisation with the lowest loss score among those compared, for all combinations of weights (w_0, w_1). For $p_0 = 0.1$ and $p_0 = 0.2$, Carsten's design is superior in almost all instances (100% of weights for $p_0 = 0.1$, 99% for $p_0 = 0.2$). For $p_0 = 0.3$, the omni-admissible design is either a Carsten design (73%) or a block design with block size two (27%). The region where the proposed design is superior is where $w_0 + w_1$ is close to one, that is, where almost all weight is on ESS(p_0, p_0) and ESS(p_0, p_1). For $p_0 = 0.4$

and $p_0 = 0.5$, the omni-admissible design is either a block design with block size two or Chen's design (85% vs 15% for $p_0 = 0.4$, 95% vs 5% for $p_0 = 0.5$). For both $p_0 = 0.4$ and $p_0 = 0.5$, the region where the proposed design is not superior to Chen's design is where $(1 - w_0 - w_1)$ is close to one, that is, where almost all weight is on N. There are no regions for which the omni-admissible design belongs to Jung's design [46], as the approach of Chen et al. [16] can be considered to produce design realisations encompassing all possible Jung design realisations but with the addition of NSC. There are also no regions for which the omni-admissible design belongs to our approach with block size eight, however this may be expected as any design using blocks of size eight will generally be outperformed by the equivalent design with blocks of size two.

Figure 3.3 shows the difference in loss scores between the block design using block size two, existing designs and block size eight, again for all possible weights. The difference is taken between the design realisations with the lowest loss scores for a given weight combination, ensuring that the best design realisation for each design approach is being compared. The loss scores have no interpretation other than as a comparison between design realisations. As with Figure 3.2, the plots show that while Carsten's design [17] is superior to the block approach for low values of p_0, it performs comparatively less well as p_0 increases (top row to bottom row). This result was also found by Chen et al. [16], and can be seen particularly on the bottom row of plots in Figure 3.3, where Carsten's designs perform poorly in comparison to the other designs. The rightmost column of plots compares block size two to block size eight, and is white or near-white at all points, indicating that the difference between the designs is always close to zero in terms of loss score. The maximum difference in loss score in favour of block size two compared to block size eight is 6 across all combinations of weights, compared to 61 for block size two compared to Carsten, 25 compared to Chen et al. and 35 compared to Jung.

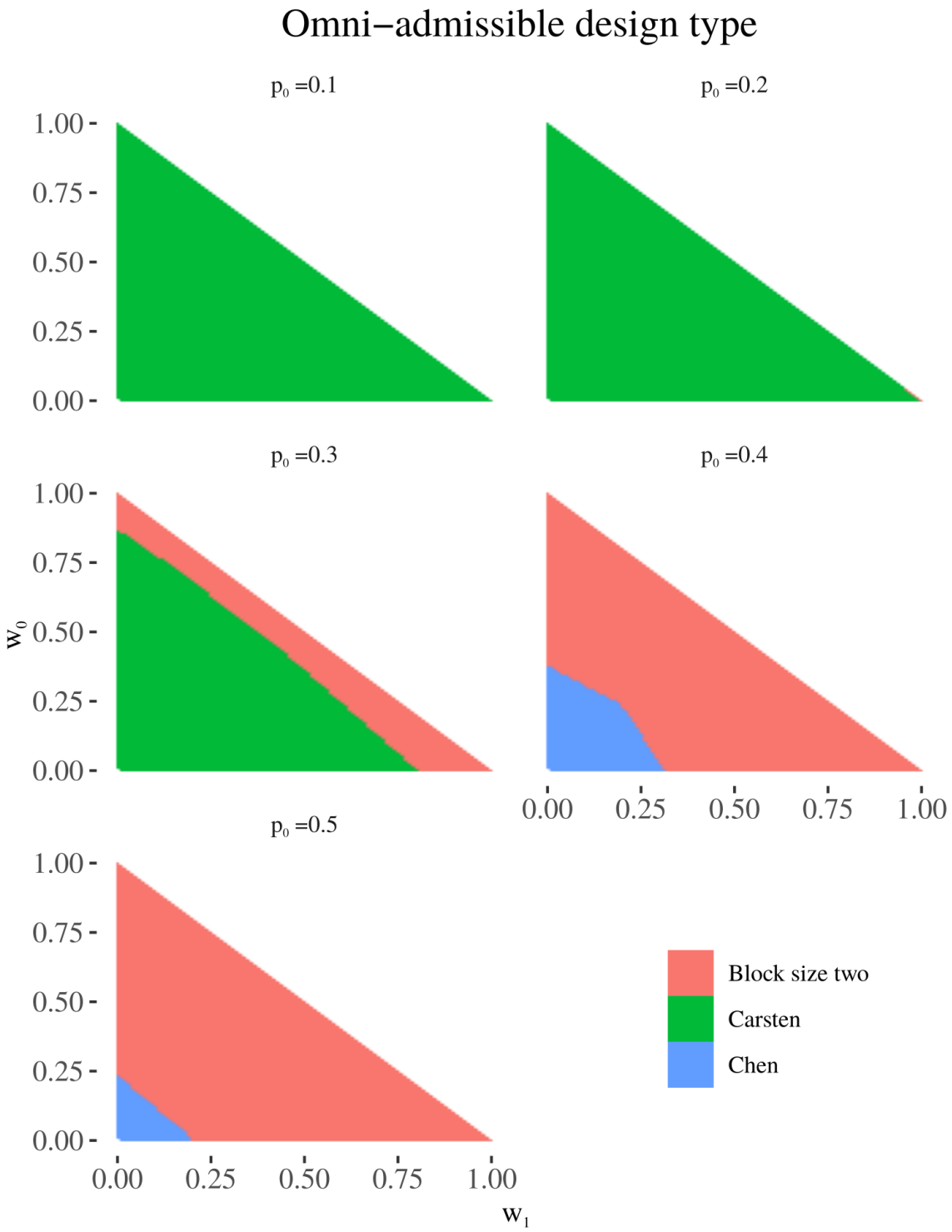


Fig. 3.2 Omni-admissible design: the approach to which the design realisation with the lowest loss score belongs, for $(\alpha,\beta) = (0.15,0.2)$, $p_0 = \{0.1,0.2,0.3,0.4,0.5\}, p_1 = p_0 + 0.2$.

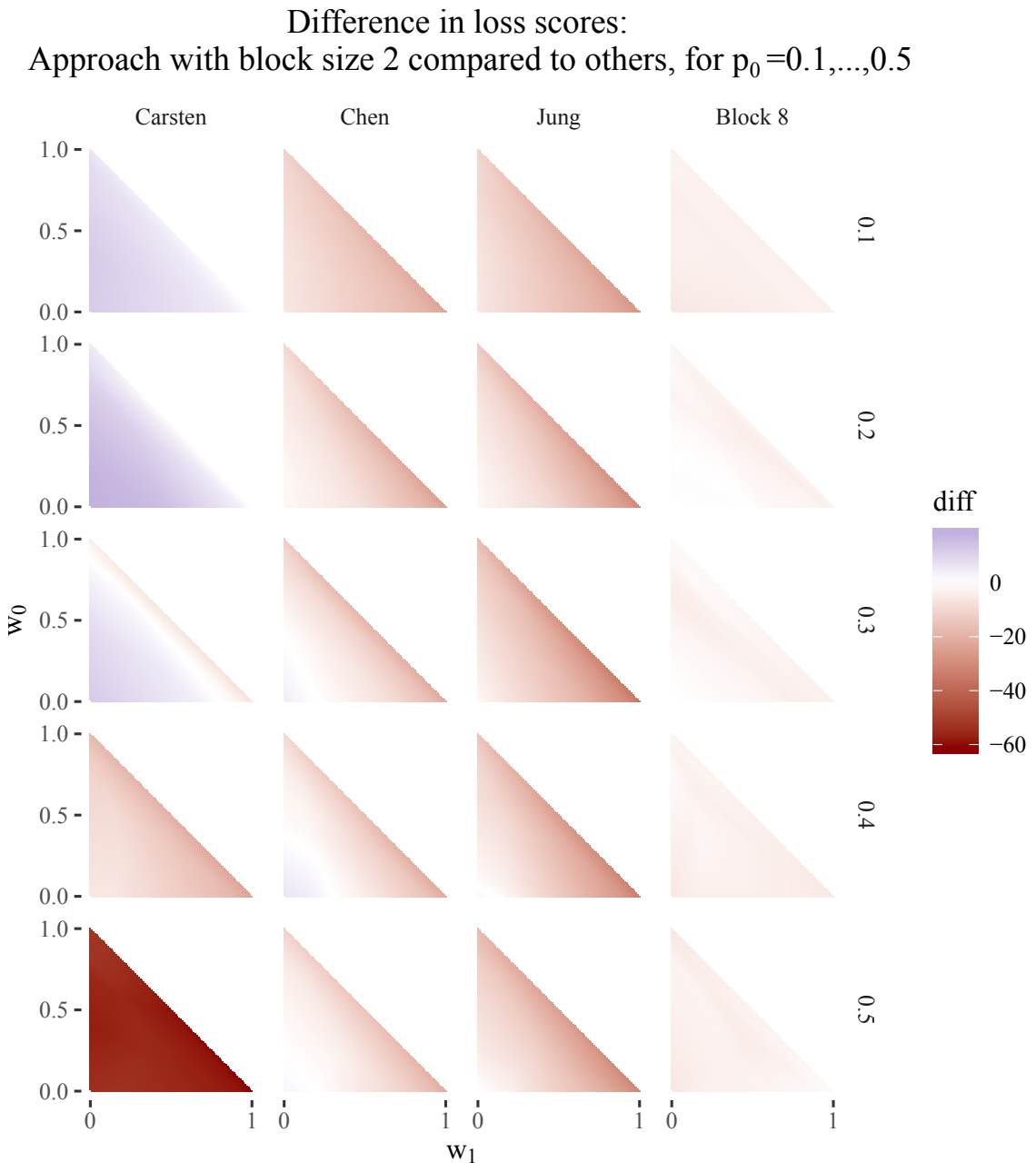


Fig. 3.3 Difference in loss scores for block design of size two versus other approaches, for $p_0 = 0.1, \ldots, 0.5$. Negative values (in red) favour the proposed design with blocks of size two.

An equivalent set of plots, comparing block size eight to existing designs, is shown in Figure 3.4, and shows similar results: both plots show the superiority of the proposed approach compared to existing designs when $p_0 \geq 0.4$, even when monitoring is reduced to conducting an interim analysis only after each block of eight participants.

Table 3.2 shows the optimal design realisations, for the two-arm designs plus Simon's design for ESS comparison, for the set of design parameters $(\alpha, \beta, p_0, p_1) = (0.15, 0.20, 0.30, 0.50)$, $p_0 = 0.3$ being the midpoint of the five values chosen for p_0. The table shows the design realisations for four optimality criteria: those that minimise $ESS(p_0, p_0)$ and $ESS(p_0, p_1)$ (the p_0- and p_1-optimal designs respectively), and those that minimise $ESS(p_0, p_0)$ and $ESS(p_0, p_1)$ among the subset of design realisations with the minimum maximum sample size N (the p_0- and p_1-minimax designs respectively). In this instance, the p_0- and p_1-minimax designs are identical for all designs considered. All designs that use curtailment are superior to Jung's design in each of the four criteria of interest (p_0- and p_1-optimal and p_0- and p_1-minimax).

For the p_0-optimal designs, the block designs achieve lower $ESS(p_0, p_0)$ than the existing randomised designs (47.3, 49.2 vs 64.9, 51.3, 60.1) at the expense of greater maximum sample size N (116, 112 vs 92, 88, 90). This is also the case for the p_1-optimal designs with regards to $ESS(p_0, p_1)$ (45.4, 49.3 vs 80.8, 52.6, 67.3 and N=112, 112 vs 82, 92, 76).

For the design parameters in Table 3.2, a standard two-arm trial with a one-sided hypothesis test and no early stopping has sample size $N = 84$, while the equivalent single-arm trial has sample size $N = 21$. As such, $ESS(p_0, p_0)$ and $ESS(p_0, p_1)$ for both block designs are closer to those of the single-arm design than the two-arm sample size under the p_0- and p_1-optimality criteria.

The cases where existing two-arm designs are superior to the proposed block designs for a single optimality criterion are the Carsten and Chen et al. designs under $p_{0/1}$-minimax, where these designs achieve a lower maximum sample size compared to the block designs

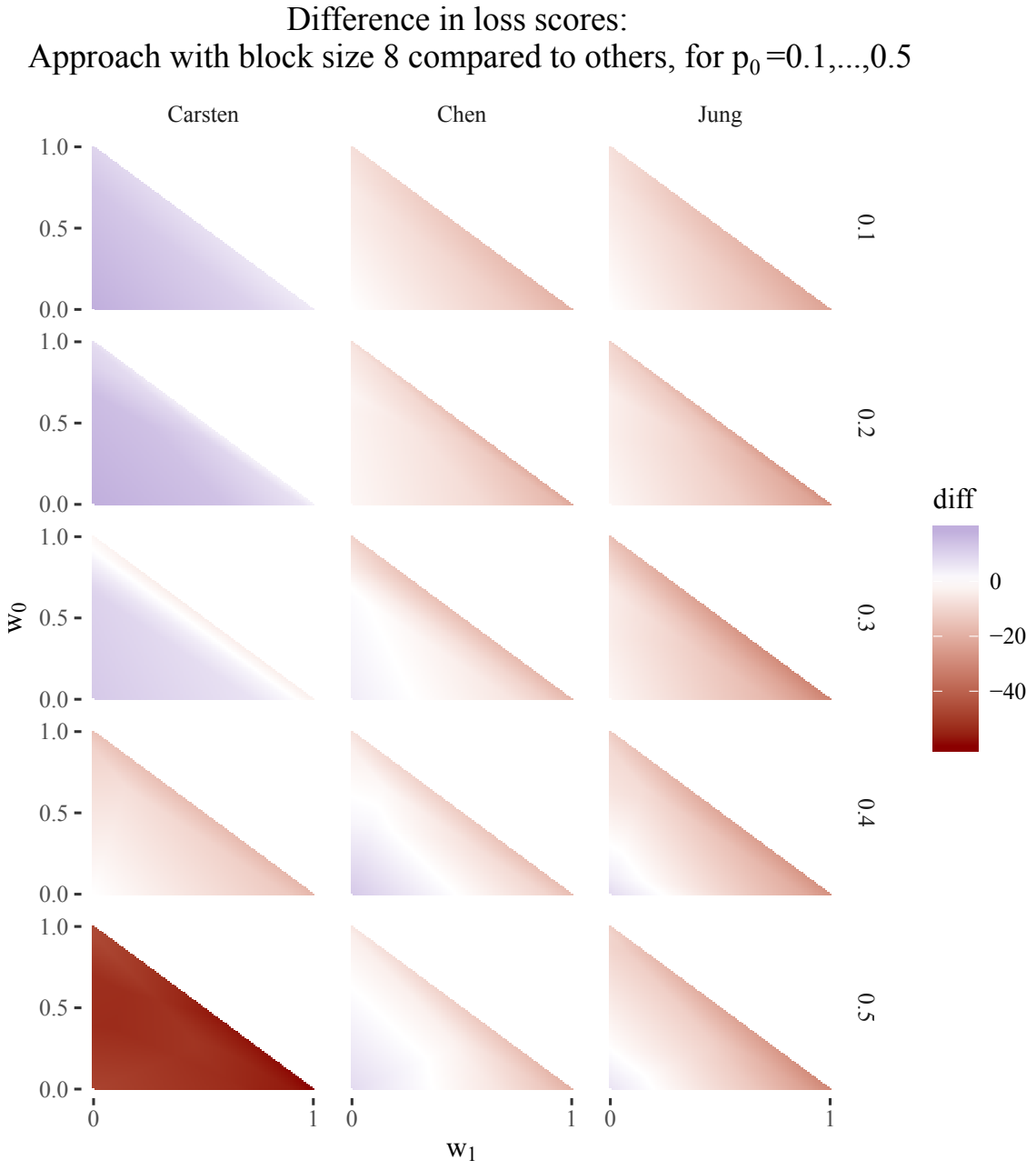


Fig. 3.4 Difference in loss scores for block design of size eight versus other approaches, for $p_0 = 0.1, \ldots, 0.5$. Negative values (in red) favour the proposed design with blocks of size eight.

	r_1	n_1	r	N_{arm}	N	$ESS(p_0, p_0)$	$ESS(p_0, p_1)$	θ_F	θ_E
p_0-optimal									
Simon	2	8	10	28	28	17.0	25.1	—	—
Jung	0	44	5	46	92	64.9	86.7	—	—
Carsten	5	38	12	44	88	51.3	60.3	0.0000	1.0000
Chen	0	32	5	45	90	60.1	76.9	0.0000	1.0000
Block 2	—	—	5	58	116	47.3	47.2	0.1348	0.9831
Block 8	—	—	5	56	112	49.2	49.3	0.3005	0.9700
p_1-optimal									
Simon	3	13	8	21	21	17.6	20.6	—	—
Jung	-1	56	5	41	82	70.5	80.8	—	—
Carsten	9	60	10	46	92	55.0	52.6	0.0000	1.0000
Chen	4	70	4	38	76	63.3	67.3	0.0000	1.0000
Block 2	—	—	6	56	112	47.9	45.4	0.1072	0.9740
Block 8	—	—	5	56	112	49.2	49.3	0.3005	0.9700
$p_{0/1}$-minimax									
Simon	3	13	8	21	21	17.6	20.6	—	—
Jung	-1	56	5	41	82	70.5	80.8	—	—
Carsten	4	40	10	34	68	53.3	52.9	0.0000	1.0000
Chen	4	70	4	38	76	63.3	67.3	0.0000	1.0000
Block 2	—	—	4	40	80	57.3	52.7	0.0428	0.9842
Block 8	—	—	4	40	80	62.2	57.1	0.0609	0.9752

Table 3.2 p_0-optimal, p_1-optimal and $p_{0/1}$-minimax designs, for $(\alpha, \beta, p_0, p_1)$=$(0.15, 0.20, 0.30, 0.50)$. N_{arm}: number of participants per arm.

(68, 76 vs 80, 80). However, the block design with block size eight requires less monitoring than the existing designs, with a maximum of 10 decisions compared to 34 decisions for Carsten and 76 for Chen et al.[17, 16]. These results are reflected in the left-hand plots of the third row of Figure 3.3, where the triangle is red near the hypotenuse, indicating superiority of the block design when minimising ESS is of greatest value, and the triangle is blue near the lower left corner, indicating superiority of Carsten's and Chen et al.'s designs when minimising maximum sample size is of greatest value.

To address possible concerns regarding very early stopping, the minimum possible number of participants was obtained for the $p_{0/1}-$optimal and $p_{0/1}-$minimax block designs for $p_0 = \{0.1, 0.2, 0.3, 0.4, 0.5\}, p_1 = p_0 + 0.2$. Across all combinations of optimality criteria and response rates, the minimum number of participants for block size two has median $9(IQR[7.5, 10])$ and the minimum number of participants for block size eight has median $8(IQR[8, 16])$. The possibility of stopping after a small number of participants is addressed in the Discussion section.

3.2.2 Comparison to group sequential design

Our proposed design would function similarly in practice to a group sequential design with many stages. As a comparison, it is possible to find group sequential designs of up to 10 stages using the `rpact` [102] package in `R` [74]. This software was used to find a design with the maximum number of stages (10) and with the design parameters as specified as Table 3.2. The design used binding stopping rules for futility and O'Brien and Fleming type alpha and beta spending. The stopping boundaries are determined by the observed difference in response rates and are shown in Table 3.3 and Figure 3.5.

The design found using `rpact` has $ESS(p_0, p_0) = 55.1$, $ESS(p_0, p_1) = 58.4$ and $N = 97$. These results may be compared to the proposed approach using blocks of size eight (Table 3.2), which finds a p_0-optimal design with $ESS(p_0, p_0) = 49.2$ (reduction of 11%), a

Stage	1	2	3	4	5	6	7	8	9	10
No go decision	*	-0.29	-0.15	-0.07	-0.02	0.02	0.04	0.06	0.08	0.10
Go decision	*	0.66	0.45	0.33	0.26	0.21	0.18	0.15	0.13	0.10

Table 3.3 Stopping boundaries for design found using package `rpact`, in terms of difference in observed response rates. *No values returned by package for stage 1 in terms of response rate: z-values were -3.246 (no go), 4.404 (go).

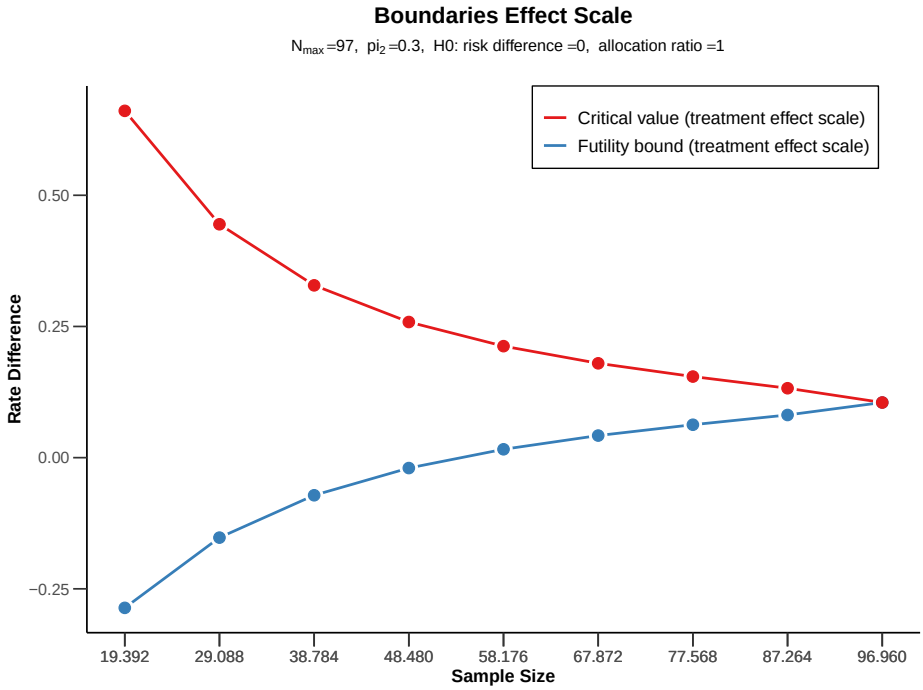


Fig. 3.5 Stopping boundaries and sample size for design found using package `rpact`, in terms of difference in observed response rates.

p_1-optimal design with $ESS(p_0, p_1) = 49.3$ (reduction of 16%) and a minimax design with $N = 80$ (reduction of 18%). The maximum number of stages in the p_0- and p_1-optimal cases would be 14 and in the minimax case would be 10, while using larger block sizes would result in fewer stages. The $p_{0/1}$-minimax design found using the proposed design with block size eight is shown in Figure 3.6. This visualisation was created using the R package `curtailment` [75], and shows at a glance the stopping boundaries for discrete numbers of participants.

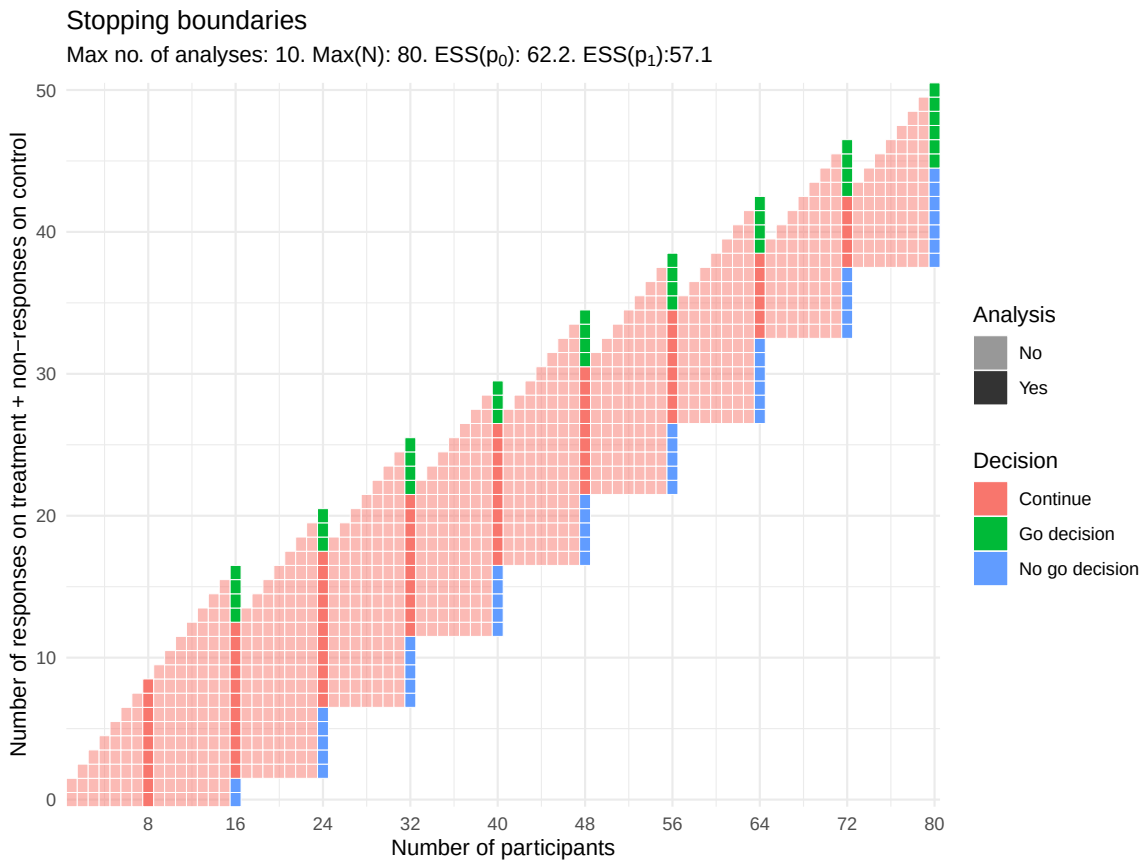


Fig. 3.6 Stopping boundaries and sample size for proposed design with block size eight from Table 3.2, $p_{0/1}$-minimax.

3.2.3 Changing true response rates

When the true response rates (p_C, p_T) differ from those specified, this can lead to a probability of rejecting H_0 that is considerably greater or lower than expected. The consequences of such deviation may depend on the design approach used.

Figure 3.7 shows the probability of rejecting H_0 when the true response rates are not equal to the specified response rates (p_0, p_1), for the p_0-optimal Carsten design and block design with block size two under $(\alpha, \beta, p_0, p_1) = (0.15, 0.20, 0.10, 0.30)$. The Carsten design has been chosen as the comparison design as it was superior to the proposed designs more often than other existing designs in Section 3.2.1. The probability of rejecting H_0 when $p_T > p_C$ is given in the lower right triangles, while the probability of rejecting H_0 when $p_T < p_C$ is given in the upper left triangles. The probability of rejecting H_0 when $p_T = p_C$ is given by the remaining diagonal.

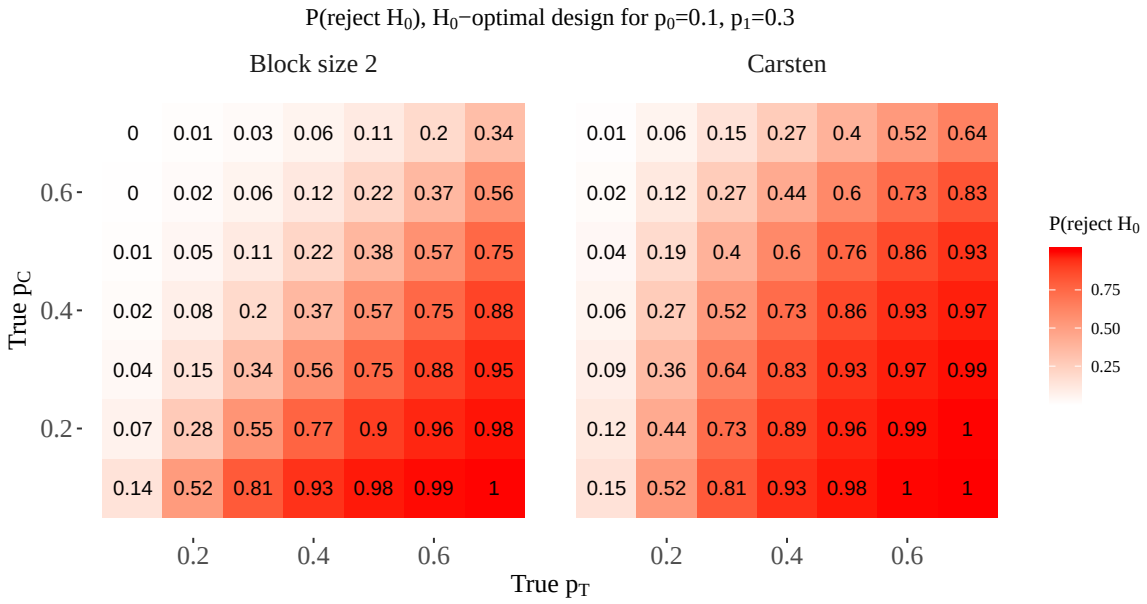


Fig. 3.7 Probability of rejecting H_0, for the p_0-optimal block size two design $(r, N, \theta_F, \theta_E) = (3, 62, 0.128, 0.932)$ and the Carsten design $(r_1, n_1, r, N) = (1, 14, 3, 64)$ under $(\alpha, \beta, p_0, p_1) = (0.15, 0.2, 0.1, 0.3)$.

Figure 3.8 shows the probability of rejecting H_0 for the p_0-optimal Carsten design and block design with block size two under $(\alpha, \beta, p_0, p_1) = (0.15, 0.20, 0.20, 0.40)$. The probabilities were obtained using simulations of size 10,000. In all instances, the probability of rejecting H_0 is greater using the Carsten design than the block design with block size two. This difference is considerable for a number of plausible pairs of response rates. For example, in Figure 3.7, where the anticipated response rates are $(p_0, p_1) = (0.1, 0.3)$, P(reject H_0)=0.64 using the Carsten design (vs 0.34 using the block design) when $(p_C, p_T) = (0.3, 0.3)$ and P(reject H_0)=0.36 (vs 0.15) when $(p_C, p_T) = (0.3, 0.2)$. Similarly, in Figure 3.8, where the anticipated response rates are $(p_0, p_1) = (0.2, 0.4)$, P(reject H_0)=0.39 (vs 0.22) when $(p_C, p_T) = (0.3, 0.3)$, and P(reject H_0)=0.51 (vs 0.25) when $(p_C, p_T) = (0.4, 0.4)$. When using Carsten's designs, if there is no difference between the treatment and control arms, and even if the treatment arm has a poorer response rate than the control arm, there may still be a substantially increased probability of rejecting H_0 and concluding that the difference in response rate is of clinical interest. This is of particular concern as a key advantage of randomised trials over single-arm trials is greater accounting for such deviations from the specified response rates [43].

3.2.4 Comparison of decision space

Given the difference in test statistics used by our design and Carsten and Chen [17], it is worthwhile to compare the decision spaces of these approaches. Using design parameters $(\alpha, \beta, p_0, p_1) = (0.15, 0.2, 0.3, 0.5)$ as per Table 3.2, in Figure 3.9 we show the decision spaces of the $p_{0/1}$-minimax design realisations for these two designs. In the block design (with block size two), the region indicating that the trial will continue is a long, thin line that narrows to a point as N is reached. This bears some resemblance to continuation region of Wald's SPRT [14], though that region does not narrow. The shape shows how stopping

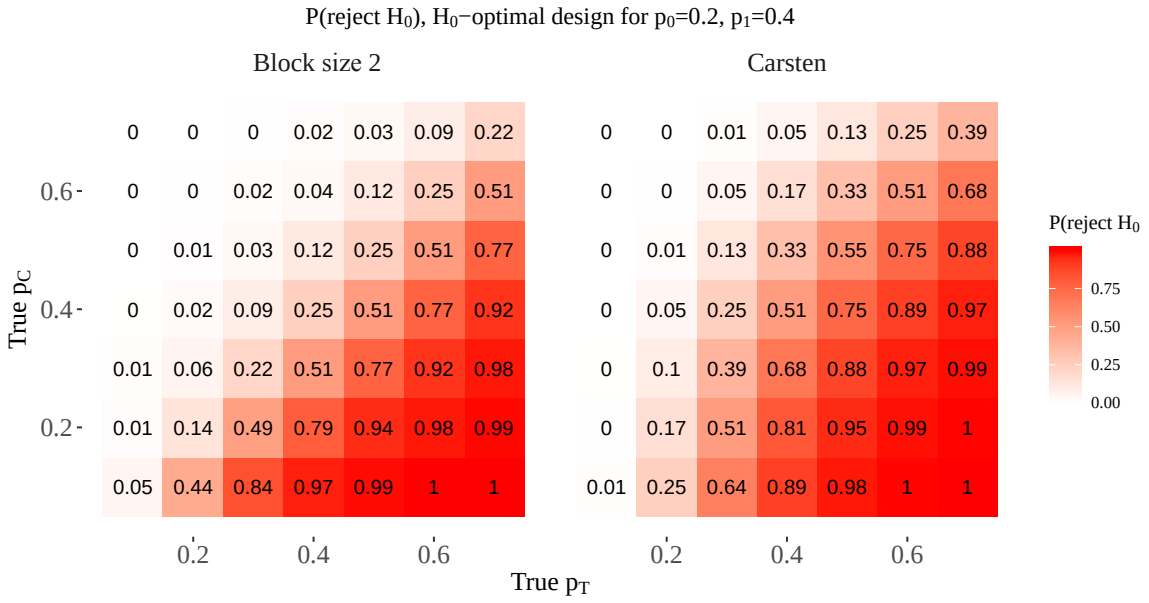
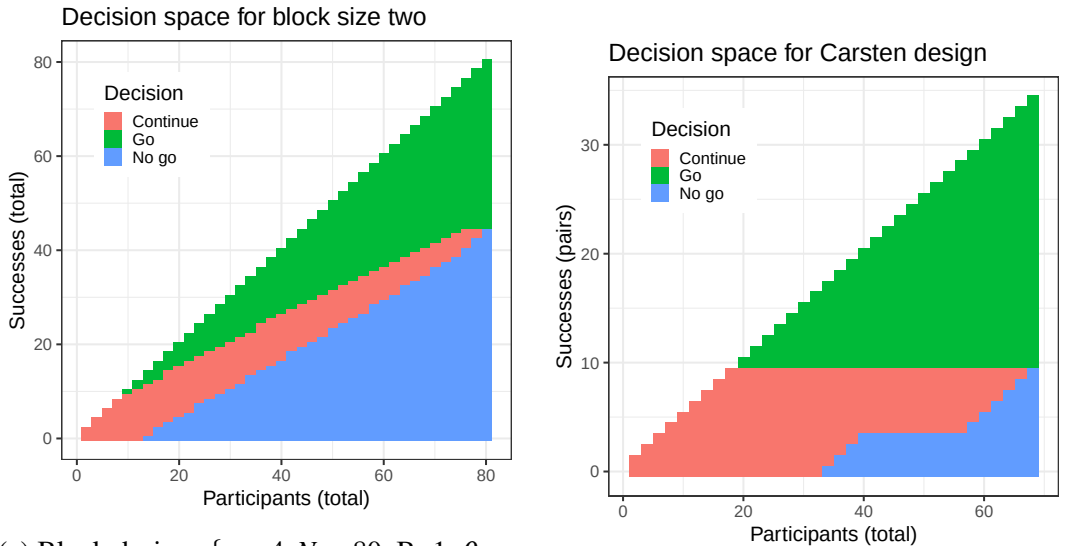


Fig. 3.8 Probability of rejecting H_0, for the p_0-optimal block size two design $(r,N,\theta_F,\theta_E) = (5,96,0.115,0.964)$ and the Carsten design $(r_1,n_1,r,N) = (2,20,7,116)$ under $(\alpha,\beta,p_0,p_1) = (0.15,0.2,0.2,0.4)$.

decisions can be made earlier than for the Carsten design, where the corresponding region for continuation is wider throughout.



(a) Block design: $\{r = 4, N = 80$, B=1, $\theta_F = 0.04276781$, $\theta_E = 0.9841904\}$.

(b) Carsten design: $\{r_1 = 4, n_1 = 40, N = 68\}$.

Fig. 3.9 Decision spaces for $p_{0/1}$-minimax design realisations, $(\alpha,\beta,p_0,p_1) = (0.15,0.2,0.3,0.5)$.

3.2.5 Real data example

A trial that has been used previously as an example in comparing two-arm binary outcome trial designs is CALGB 50502, a randomized phase II trial for the treatment of Hodgkin Lymphoma [97, 99, 101]. The design parameters of the trial are $(\alpha, \beta, p_0, p_1) = (0.15, 0.20, 0.70, 0.85)$. Optimal designs for this set of design parameters were sought for the designs of Jung, Carsten and Chen, Chen et al. [46, 17, 16] and the block designs, using the same methods as for the main comparisons. The maximum sample size searched over was 200, with the exception of the Carsten and Chen design, where the maximum sample size was 400. However, no feasible designs were found using the Carsten and Chen design. This is not surprising, as Chen et al. [16] showed that the maximum and expected sample size of the Carsten and Chen design increases rapidly with p_0, reaching $N = 278$, ESS(p_0, p_0)=162 for design parameters $(\alpha = 0.05, \beta = 0.1, p_0 = 0.6, p_1 = 0.9)$. Table 3.4 shows the p_0- and p_1-optimal and p_0- and p_1-minimax designs for the remaining designs. The p_0-minimax and p_1-minimax designs were again identical. The p_0- and p_1-optimal block designs reduce ESS by approximately one third compared to the existing designs, at the expense of increased maximum sample size. The maximum sample size for the $p_{0/1}$-minimax designs are similar across all four two-arm designs, in the range $[122, 128]$, though here $ESS(p_0, p_0)$ and $ESS(p_0, p_1)$ are superior for the block designs compared to the existing designs.

For the design parameters in this example, a standard two-arm trial with a one-sided hypothesis test and no early stopping has sample size $N = 108$, while the equivalent single has sample size $N = 31$. As such, $ESS(p_0, p_0)$ for both block designs are closer to the single-stage single-arm sample size than the two-arm sample size under the p_0- and p_1-optimality criteria.

	r_1	n_1	r	N_{arm}	N	$ESS(p_0,p_0)$	$ESS(p_0,p_1)$	θ_F	θ_E
Single-stage	—	—	—	54	108	108	108	—	—
p_0-optimal									
Simon	10	14	25	33	33	20.7	30.2	—	—
Jung	-1	54	6	73	146	94.6	135.1	—	—
Chen	0	46	6	72	144	93.8	124.6	0.0000	1.0000
Block 2	—	—	6	99	198	61.1	79.4	0.1108	0.9928
Block 8	—	—	4	88	176	64.4	87.7	0.3391	0.9965
p_1-optimal									
Simon	4	7	23	30	30	21.9	28.3	—	—
Jung	3	112	5	62	124	114.1	121.8	—	—
Chen	1	98	6	61	122	102.9	113.2	0.0000	1.0000
Block 2	—	—	6	99	198	61.1	79.4	0.1108	0.9928
Block 8	—	—	6	92	184	66.4	83.5	0.2730	0.9866
$p_{0/1}$-minimax									
Simon	20	26	22	29	29	26.5	28.4	—	—
Jung	3	112	5	62	124	114.1	121.8	—	—
Chen	0	92	6	61	122	102.1	113.4	0.0000	1.0000
Block 2	—	—	5	62	124	96.4	95.9	0.0064	0.9960
Block 8	—	—	5	64	128	80.1	91.7	0.1304	0.9887

Table 3.4 p_0-optimal, p_1-optimal and $p_{0/1}$-minimax designs, for (α,β,p_0,p_1)=$(0.15,0.20,0.70,0.85)$. N_{arm}: number of participants per arm.

3.3 Discussion

This chapter introduces a new design for two-arm phase II binary outcome clinical trials. While sequential monitoring has previously been used in conjunction with NSC, this design is novel as it uses SC to reduce ESS. Curtailment may occur due to observing either a high or low response rate on the treatment arm compared to the control arm. Participants are allocated in randomised blocks, and trial results are noted after each block and compared to specified stopping boundaries. The trial will end if the required final difference between the arm response rates is either certain to be reached or certain to not be reached. Additionally, the trial will end if the CP is either greater than some upper threshold θ_E or less than some lower threshold, θ_F. These thresholds, in combination with the maximum sample size N, the required final difference in treatment arm response rates r and desired type-I error-rate and power determine the stopping boundaries.

The probability of rejecting the null hypothesis is controlled be at most α when $p_C = p_T = p_0$ and at least $1-\beta$ when $p_C = p_0, p_T = p_1$. However, if the true response rates differ from the specified response rates, the probability of rejecting the null hypothesis may be affected. This has been addressed in Section 3.2.3. For the proposed designs, the type-I error-rate is maximised at $p_C = p_T = 0.5$, and so this error rate could be controlled over the interval $[0,1]$ by setting $p_0 = 0.5$. However, this choice may not accurately reflect an investigator's belief regarding the anticipated response rates.

The proposed block design was compared to three existing designs, described in Jung, Carsten and Chen and Chen et al. [46, 17, 16]. All three designs include an interim analysis, while the designs of Carsten and Chen and Chen et al. also use NSC. A comparison between the proposed design and the three existing designs was undertaken using a loss function, a weighted sum of three optimality criteria. The type-I error-rate was set to $\alpha = 0.15$ and power to $1-\beta = 0.8$, as in Table 1 of Jung [46]. Five sets of response rates (p_0, p_1) were examined. For low values of p_0, only the Carsten and Chen design was superior to the proposed block

design. The superior performance of the Carsten and Chen design for low values of p_0 has been previously noted by Chen et al. [16]. However, not discussed by Carsten and Chen nor Chen et al. is sensitivity of the Carsten and Chen design to deviations from the specified response rates. Such deviations can lead to a considerable increase in the probability of rejecting H_0 when the treatment is not sufficiently superior to control. For greater values of p_0, the block design is superior to the compared designs for most combinations of weights, in terms of expected and maximum sample size, even when using blocks of size eight. Under the given requirements for type-I error-rate and power, the ESS of the design with block size eight is likely to be less than or approximately equal to that obtained using the designs of Carsten and Chen or Chen et al. for $p_0 \geq 0.3$, and with the degree of monitoring reduced by a factor of four or eight respectively.

The designs were also compared using a real-life example, used previously to compare two-arm designs [46, 99, 101]. When minimising ESS under either $p_C = p_T = p_0$ or $p_C = p_0, p_T = p_1$, the reduction in ESS for the proposed block designs was considerable compared to existing designs. When minimising maximum sample size, the proposed block designs had comparable maximum sample size and smaller ESS compared to existing designs, again with monitoring frequency reduced considerably when using blocks of size eight.

The designs of Carsten and Chen and Chen et al. [17, 16] are examples of continuous monitoring, where, in contrast to the two-stage designs of Simon [5] and Jung [46], the data are subject to more frequent interim analyses. When continuous monitoring is used in a clinical trial, the actual sample size is dependent on the number of participants' responses available at each interim analysis. Monitoring may take place after every participant or less frequently [21]. Continuous monitoring is of greatest value when endpoint length is short, for example if, in oncology, tumour response is measured over short periods of time, though it is possible to use curtailment and continuous monitoring for endpoint lengths that may be considered long [32]. Given the low recruitment rate of randomised controlled

trials, for example, the median rate of 0.92 participants per centre per month reported in a review by Walters et al. [20], the effect of any lag on observed sample size is likely to be small. Furthermore, trial recruitment rates are generally lower than expected, favouring more frequent monitoring [19].

The designs of Jung, Carsten and Chen and Chen et al. [46, 17, 16] use what are known as binding stopping rules, whereby stopping is mandatory when any pre-specified stopping boundary is reached. This is in contrast to non-binding stopping rules, where, despite reaching a stopping boundary, a trial may continue for other reasons, for example, to gain more information regarding adverse events [103]. Despite binding stopping rules being present, some trials have disregarded the planned stopping rules in practice using the same rationale as NSC, that is, the final decision is known with certainty due to the results so far. Such curtailment has been used both due to low and high observed response rates, and can only have been done by reviewing the results frequently. Numerous examples of this were described in Section 1.1.3.1. As such, continuous monitoring is being used in some trials where none is specified.

An advantage of using the block design over existing curtailed designs is that fewer interim analyses may be required. While Carsten and Chen's design [17] requires monitoring after every pair of participants and Chen et al.'s design [16] after every single participant, the degree of monitoring required for the block design depends on the block size used, and may be specified by the investigator. Furthermore, use of larger blocks reduces computational burden with regards to the search for design realisations, with only a small increase in ESS.

This chapter shows the benefit of using the proposed approach, which combines SC, randomised blocks and other features in a novel way. It provides the exact distribution of a trial's outcomes, meaning that its operating characteristics are known without sampling error. Compared to other existing two-arm designs, the proposed approach considerably reduces ESS.

Chapter 4

Multi-outcome trials with a generalised number of efficacious outcomes

4.1 Brief description of existing multi-outcome multi-stage designs

The main limitation of existing work is that current multi-outcome multi-stage designs focus almost entirely on evaluating if all outcomes show evidence of efficacy or if at least one outcome shows evidence of efficacy. While Delorme et al. [72] and Mielke et al. [73] provide multi-outcome designs that evaluate when a general number of outcomes show promise, these designs are single-stage only. Using a single-stage design means that there are no interim analyses and no decisions made until the end of trial. In single-stage multi-outcome trials, the sample size is fixed and every outcome is measured for every participant. We propose two designs that provide this design characteristic in a multi-stage setting. Beyond this, many multi-outcome multi-stage designs allow only a maximum of two outcomes, while the proposed designs permit any number of outcomes. Finally, one of the two proposed designs permits ceasing measurement of an outcome that is performing poorly. While this

design characteristic is not novel on its own, we believe that this property has not been implemented in a design that evaluates multiple outcomes powered under the condition of a general number of outcomes showing promise.

4.2 Proposed designs

In both proposed designs, we subsume the concepts of co-primary and multiple primary outcomes into a general framework of single-arm designs that permit rejection of a null hypothesis H_0 when promising effects are observed on some specified m out of K outcomes. We apply this concept to multi-outcome multi-stage design, allowing the trial to end at any stage, for either a go decision (reject H_0) or a no go decision (do not reject H_0). The first proposed design permits any number of stages J. This design will be compared to a multi-stage composite design, where again the trial may end at any stage, for a go or no go decision, and a single, composite outcome is evaluated at each stage. The second proposed design limits the number of stages to two, and permits dropping poorly-performing outcomes at the interim analysis while still allowing the trial to end at this point for a go decision or no go decision. This design is compared to a multi-outcome single-stage design that, like both proposed designs, rejects the null hypothesis when promising effects are observed on m out of K outcomes.

4.3 Methods: Multi-outcome multi-stage design with general number of required efficacious outcomes

Let K be the total number of (continuous) outcomes that will be measured in the trial. Let J be the maximum number of allowed stages of the design. The number of participants in each stage of the trial is denoted by n. The maximum sample size is then $N = Jn$. We let

$X_{ik}, i = 1,\ldots,Jn, k = 1,\ldots,K$ be the response in participant i for outcome k. The responses are assumed to have the following multivariate normal distribution:

$$\begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{iK} \end{pmatrix} \sim MVN_K \left[\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \ldots & \rho_{1K}\sigma_1\sigma_K \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \ldots & \rho_{2K}\sigma_2\sigma_K \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{K1}\sigma_K\sigma_1 & \rho_{K2}\sigma_K\sigma_2 & \ldots & \sigma_K^2 \end{pmatrix} \right].$$

As noted, we assume that interest lies in whether m or more outcomes show promise. Using a single hypothesis approach, the null and alternative hypotheses are

$$H_0 : \sum_{k=1}^{K} \mathbb{I}(\mu_k > 0) < m, \qquad H_1 : \sum_{k=1}^{K} \mathbb{I}(\mu_k > 0) \geq m. \tag{4.1}$$

After each stage j, an interim analysis is undertaken at which point the trial may stop for either a go decision or no go decision. The lower and upper stopping boundaries at stage j are denoted f_j and e_j respectively. The test statistic for outcome k at stage j is $Z_{jk} = \hat{\tau}_{jk}\sqrt{\mathscr{I}_j} = \hat{\tau}_{jk}\sqrt{jn/\sigma_k^2}$, where $\hat{\tau}_{jk} = \sum_{i=1}^{jn} x_{ik}/jn$ is the observed effect for outcome k at analysis j. The trial will end and the null hypothesis will be rejected if m of the test statistics Z_{jk} simultaneously exceed upper stopping boundary e_j, i.e. if

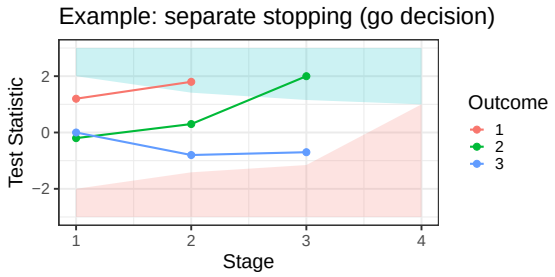
$$\sum_{k=1}^{K} \mathbb{I}(Z_{jk} > e_j) \geq m, \text{for any } j.$$

Conversely, a trial will end and the null hypothesis will not be rejected if $K - m + 1$ outcomes are simultaneously lower than lower stopping boundary f_j, i.e. if

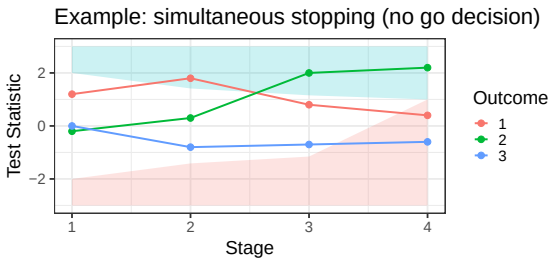
$$\sum_{k=1}^{K} \mathbb{I}(Z_{jk} < f_j) \geq K - m + 1, \text{for any } j.$$

This is a *simultaneous* multi-stage approach. This is in contrast to a *separate* multi-stage approach, where there are K separate hypotheses, one for each outcome, each of which may be rejected (or not) independently of one another [57, 68]. In a separate approach, a decision to reject or not reject some hypothesis $H_k, k = 1,\ldots,K$ is permitted at any stage, and occurs when the corresponding test statistic crosses an upper or lower stopping boundary. Once a decision has been made regarding H_k, measurement of outcome k will end. This reduces the number of outcome measurements made. Reducing the expected number of measurements (ENM) may be of particular interest if there are some outcomes that we desire to minimise, either due to cost or otherwise [104], and using the separate multi-stage approach is one way of doing this. A visual comparison is provided in Figure 4.1, where we show an example design with $J = 4$ stages, $K = 3$ outcomes and number of outcomes required to show promise $m = 2$. In this example, and for the first proposed design, we use stopping boundaries of the form proposed by Wang and Tsiatis [105], for which the stopping boundaries can be characterised by scalars C and Δ: $e_j = Cj^{\Delta-0.5}$, $j = 1,\ldots,J$. Similarly $f_j = -Cj^{\Delta-0.5}$ for $j = 1,\ldots,J-1$ and $f_j = e_J$ for $j = J$ to ensure a decision is reached by the final stage. This is a generalisation of the boundaries proposed by Pocock [106] and by O'Brien and Fleming [107], which are special cases equivalent to $\Delta = 0.5$ and $\Delta = 0$ respectively. Outcome-specific boundaries $C_k, k = 1,\ldots,K$, could theoretically be obtained through K-dimensional optimisation. However, given the definition of type-I error-rate used, detailed in Section 4.3.2, there are potentially infinite sets of K constants that satisfy any required type-I error-rate. Therefore, we do not consider this possibility further. Figure 4.1a shows the separate approach: at stage 2, outcome 1 crosses the upper boundary and is no longer measured; at stage 3, a second outcome cross the upper boundary, meaning that $m = 2$ outcomes have separately shown promise, and the trial ends. In Figures 4.1b

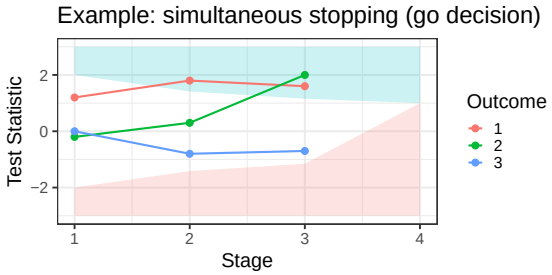
and 4.1c, the same initial data are shown, but in a simultaneous approach. Now, having a single outcome cross the upper boundary at stage 2 (or any stage) has no effect on the subsequent number of outcomes measured – all outcomes continue to be measured until either m outcomes simultaneously cross the upper boundary or $K-m+1=2$ outcomes cross the lower boundary. In the example in Figure 4.1b, outcome 1 crosses back over the upper boundary at stage 3, and the trial continues. At stage 4, $K-m+1$ outcomes simultaneously cross the lower boundary and consequently a no go decision is made. In the example in Figure 4.1c, outcome 1 remains above the upper boundary at stage 3, and so m outcomes have simultaneously crossed the upper boundary, and consequently a go decision is made.



(a) Example of separate stopping approach. Go decision at stage 3.



(b) Example of simultaneous stopping approach. No go decision at stage 4.



(c) Example of simultaneous stopping approach. Go decision at stage 3.

Fig. 4.1 Examples of separate and simultaneous stopping approaches, $K=3, m=2$.

4.3.1 Covariance structure

The covariance structure must be derived for the multivariate normal distribution of the test statistics across differing stages and outcomes. Although the proposed designs have sample

size jn at stage j, that is, with an equal number of participants at each stage, for the sake of generality we derive the covariance matrix for a general number of participants at each stage. Define n_j, $j = 1,\ldots,J$ to be the sample size at stage j, and N_j to be the total sample size at stage j, that is, $N_j = n_1 + n_2 + \cdots + n_j = \sum_{i=1}^{j} n_i$. For a single outcome k, the covariance of two test statistics at stages j_A, j_B, $j_B \geq j_A$ is

$$\begin{aligned}
\operatorname{cov}(Z_{j_A k}, Z_{j_B k}) &= \operatorname{cov}\left(\sqrt{\frac{N_{j_A}}{\sigma_k^2}}\hat{\mu}_{j_A k}, \sqrt{\frac{N_{j_B}}{\sigma_k^2}}\hat{\mu}_{j_B k}\right) \\
&= \sqrt{\frac{N_{j_A}}{\sigma_k^2}}\sqrt{\frac{N_{j_B}}{\sigma_k^2}}\operatorname{cov}\left(\hat{\mu}_{j_A k}, \hat{\mu}_{j_B k}\right) \\
&= \sqrt{\frac{N_{j_A}}{\sigma_k^2}}\sqrt{\frac{N_{j_B}}{\sigma_k^2}}\operatorname{cov}\left(\frac{1}{N_{j_A}}\sum_{i=1}^{N_{j_A}} X_{ik}, \frac{1}{N_{j_B}}\sum_{i=1}^{N_{j_B}} X_{ik}\right) \\
&= \sqrt{\frac{N_{j_A}}{\sigma_k^2}}\sqrt{\frac{N_{j_B}}{\sigma_k^2}}\frac{1}{N_{j_A}}\frac{1}{N_{j_B}}\operatorname{cov}\left(\sum_{i=1}^{N_{j_A}} X_{ik}, \sum_{i=1}^{N_{j_B}} X_{ik}\right) \\
&= \frac{1}{\sigma_k^2}\sqrt{\frac{1}{N_{j_A}}}\sqrt{\frac{1}{N_{j_B}}}\sum_{i=1}^{N_{j_A}}\operatorname{cov}\left(X_{ik}, X_{ik}\right) \\
&= \frac{1}{\sigma_k^2}\sqrt{\frac{1}{N_{j_A}}}\sqrt{\frac{1}{N_{j_B}}}N_{j_A}\sigma_k^2 \\
&= \sqrt{\frac{N_{j_A}}{N_{j_B}}}
\end{aligned} \tag{4.2}$$

For a single stage j, the correlation coefficient between two test statistics for outcomes k_1, k_2, $k_1 \neq k_2$ is $\rho_{k_1 k_2}$. The covariance $\operatorname{cov}(Z_{jk_1}, Z_{jk_2})$ is then

$$
\begin{aligned}
\operatorname{cov}(Z_{jk_1}, Z_{jk_2}) &= \operatorname{cov}\left(\frac{\hat{\mu}_{k_1}}{\sqrt{\sigma_{k_1}^2/n_j}}, \frac{\hat{\mu}_{k_2}}{\sqrt{\sigma_{k_2}^2/n_j}}\right) \\
&= \sqrt{\frac{n_j}{\sigma_{k_1}^2}}\sqrt{\frac{n_j}{\sigma_{k_2}^2}}\operatorname{cov}(\hat{\mu}_{k_1}, \hat{\mu}_{k_2}) \\
&= \frac{n_j}{\sqrt{\sigma_{k_1}^2\sigma_{k_2}^2}}\operatorname{cov}\left(\frac{1}{n_j}\sum_{i=1}^{n_j} X_{ik_1}, \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ik_2}\right) \\
&= \frac{1}{n_j\sqrt{\sigma_{k_1}^2\sigma_{k_2}^2}}\operatorname{cov}\left(\sum_{i=1}^{n_j} X_{ik_1}, \sum_{i=1}^{n_j} X_{ik_2}\right) \\
&= \frac{1}{n_j\sqrt{\sigma_{k_1}^2\sigma_{k_2}^2}}\sum_{i=1}^{n_j}\operatorname{cov}(X_{ik_1}, X_{ik_2}) \\
&= \frac{1}{n_j\sqrt{\sigma_{k_1}^2\sigma_{k_2}^2}} n_j \rho_{k_1k_2}\sigma_{k_1}\sigma_{k_2} \\
&= \rho_{k_1k_2} \qquad (4.3)
\end{aligned}
$$

The covariance of two test statistics for stages j_A, j_B, $j_B \geq j_A$ and outcomes k_1, k_2, $k_1 \neq k_2$, is

$$
\begin{aligned}
\operatorname{cov}\left(Z_{j_A k_1}, Z_{j_B k_2}\right) &= \operatorname{cov}\left(\sqrt{\frac{N_{j_A}}{\sigma_{k_1}^2}}\hat{\mu}_{j_A k_1}, \sqrt{\frac{N_{j_B}}{\sigma_{k_2}^2}}\hat{\mu}_{j_B k_2}\right) \\
&= \sqrt{\frac{N_{j_A}}{\sigma_{k_1}^2}}\sqrt{\frac{N_{j_B}}{\sigma_{k_2}^2}}\operatorname{cov}\left(\hat{\mu}_{j_A k_1}, \hat{\mu}_{j_B k_2}\right) \\
&= \sqrt{\frac{N_{j_A}}{\sigma_{k_1}^2}}\sqrt{\frac{N_{j_B}}{\sigma_{k_2}^2}}\operatorname{cov}\left(\frac{1}{N_{j_A}}\sum_{i=1}^{N_{j_A}} X_{ik_1}, \frac{1}{N_{j_B}}\sum_{i=1}^{N_{j_B}} X_{ik_2}\right) \\
&= \sqrt{\frac{N_{j_A}}{\sigma_{k_1}^2}}\sqrt{\frac{N_{j_B}}{\sigma_{k_2}^2}}\frac{1}{N_{j_A}}\frac{1}{N_{j_B}}\operatorname{cov}\left(\sum_{i=1}^{N_{j_A}} X_{ik_1}, \sum_{i=1}^{N_{j_B}} X_{ik_2}\right) \\
&= \frac{1}{\sqrt{\sigma_{k_1}^2\sigma_{k_2}^2 N_{j_A} N_{j_B}}}\sum_{i=1}^{N_{j_A}}\operatorname{cov}\left(X_{ik_1}, X_{ik_2}\right) \\
&= \frac{1}{\sqrt{\sigma_{k_1}^2\sigma_{k_2}^2 N_{j_A} N_{j_B}}} N_{j_A}\rho_{k_1k_2}\sigma_{k_1}\sigma_{k_2} \\
&= \sqrt{\frac{N_{j_A}}{N_{j_B}}}\rho_{k_1k_2} \qquad (4.4)
\end{aligned}
$$

Combining Equations (4.2), (4.3) and (4.4), the covariance $\operatorname{cov}(Z_{j_A k_1}, Z_{j_B k_2})$ for any $j_A, j_B,\ j_B \geq j_A$ and k_1, k_2 can be stated as

$$
\operatorname{cov}(Z_{j_A k_1}, Z_{j_B k_2}) = \begin{cases} 1 & \text{if } j_A = j_B \text{ and } k_1 = k_2 \\ \rho_{k_1k_2} & \text{if } j_A = j_B \text{ and } k_1 \neq k_2 \\ \sqrt{\frac{N_{j_A}}{N_{j_B}}} & \text{if } j_A \neq j_B \text{ and } k_1 = k_2 \\ \rho_{k_1k_2}\sqrt{\frac{N_{j_A}}{N_{j_B}}} & \text{if } j_A \neq j_B \text{ and } k_1 \neq k_2 \end{cases}. \qquad (4.5)
$$

This allows the construction of a covariance matrix for test statistics, for any number of stages J and outcomes K. This covariance matrix is necessary to describe the multivariate normal distribution of the test statistics, shown in Equation (4.6) directly below. Note: in this

equation, each element of the covariance matrix $\text{cov}(Z_{jk}, Z_{jk})$ is presented simply as jk, jk to save space.

$$\begin{pmatrix} Z_{j_1k_1} \\ Z_{j_1k_2} \\ \vdots \\ Z_{j_1K} \\ Z_{j_2k_1} \\ Z_{j_2k_2} \\ \vdots \\ \vdots \\ Z_{JK} \end{pmatrix} \sim MVN_{JK} \left(\begin{pmatrix} \hat{\tau}_{j_1k_1}\sqrt{\frac{N_{j_1}}{\sigma^2_{k_1}}} \\ \hat{\tau}_{j_1k_2}\sqrt{\frac{N_{j_1}}{\sigma^2_{k_2}}} \\ \vdots \\ \hat{\tau}_{j_1K}\sqrt{\frac{N_{j_1}}{\sigma^2_{K}}} \\ \hat{\tau}_{j_2k_1}\sqrt{\frac{N_{j_2}}{\sigma^2_{k_1}}} \\ \hat{\tau}_{j_2k_2}\sqrt{\frac{N_{j_2}}{\sigma^2_{k_2}}} \\ \vdots \\ \vdots \\ \hat{\tau}_{JK}\sqrt{\frac{N_{J}}{\sigma^2_{K}}} \end{pmatrix}, \begin{pmatrix} j_1k_1, j_1k_1 & j_1k_1, j_1k_2 & \dots & j_1k_1, j_1K & j_1k_1, j_2k_1 & j_1k_1, j_2k_2 & \dots & \dots & j_1k_1, JK \\ j_1k_2, j_1k_1 & j_1k_2, j_1k_2 & \dots & j_1k_2, j_1K & j_1k_2, j_2k_1 & j_1k_2, j_2k_2 & \dots & \dots & j_1k_2, JK \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ j_1K, j_1k_1 & j_1K, j_1k_2 & \dots & j_1K, j_1K & j_1K, j_2k_1 & j_1K, j_2k_2 & \dots & \dots & j_1K, JK \\ j_2k_1, j_1k_1 & j_2k_1, j_1k_2 & \dots & j_2k_1, j_1K & j_2k_1, j_2k_1 & j_2k_1, j_2k_2 & \dots & \dots & j_2k_1, JK \\ j_2k_2, j_1k_1 & j_2k_2, j_1k_2 & \dots & j_2k_2, j_1K & j_2k_2, j_2k_1 & j_2k_2, j_2k_2 & \dots & \dots & j_2k_2, JK \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ JK, j_1k_1 & JK, j_1k_2 & \dots & JK, j_1k_1 & JK, j_1K & JK, j_2k_1 & \dots & \dots & JK, JK \end{pmatrix} \right) \tag{4.6}$$

4.3.2 Type-I error-rate and power

Define $R(\boldsymbol{\mu}|K,m,J,C,\Delta)$ as the probability of rejecting the null hypothesis when the true outcome effects are equal to $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$, for some design realisation characterised by K,m,J,C and Δ. The probability $R(\boldsymbol{\mu}|K,m,J,C,\Delta)$ can be readily evaluated using simulation (described in Section 4.3.3). We define type-I error-rate as

$$\alpha^* = R(\boldsymbol{\mu} = \mathbf{0}|K,m,J,C,\Delta).$$

That is, we control the type-I error-rate under the scenario where $\boldsymbol{\mu} = \mathbf{0}$. Cook and Farewell [58] have previously used this manner of type-I error control in a multiple outcome setting. This is in contrast to Lehmann and Romano [108], who treat each hypothesis separately and describe controlling the probability of rejecting k true hypotheses as the k-familywise error rate (where $k \equiv m$ here). Dmitrienko et al. [109] refer to this as the generalized familywise error rate while Grayling et al. [71] describe this as the a-generalised type-I familywise error rate. The familywise error-rate is the probability of rejecting at least one true null hypothesis, with the understanding that this error-rate increases as the number of hypotheses increases. Our focus is on the probability of making a certain decision, specifically, of rejecting the null hypothesis, as the design is framed using a single null hypothesis, rather than a separate null hypothesis for each outcome. This makes direct comparison with weak and strong control of the familywise error-rate difficult. While weak control of the familywise error-rate at some level α ensures that the familywise error-rate is less than or equal to α when all null hypotheses are true and strong control ensures this for all configurations of null hypothesis, our design ensures that the probability of concluding that at least m outcomes are efficacious when all outcomes have effect size zero is less than or equal to α. In the absence of separate hypotheses, this definition of type-I error-rate controls addresses the same underlying issue as familywise error-rate, that is, making an

incorrect conclusion. Weak control is somewhat analogous to controlling type-I error-rate under a specific set of outcome effects (for example, $\boldsymbol{\mu} = \mathbf{0}$) while strong control is somewhat analogous to controlling type-I error-rate at every set of outcome effects such that H_0 is true.

We define power as

$$1 - \beta^* = R\left(\boldsymbol{\mu} = \boldsymbol{\delta}_\beta | K, m, J, C, \Delta\right),$$

for some vector of effect sizes $\boldsymbol{\delta}_\beta = (\delta_{\beta 1}, \delta_{\beta 2}, \ldots, \delta_{\beta K})$ for which we would like to control the probability of rejecting H_0.

Let the required type I and type II errors be α and β. We require designs that satisfy the conditions $\alpha^* = R(\boldsymbol{\mu} = \mathbf{0} | K, m, J, C, \Delta) \leq \alpha$ and $1 - \beta^* = R\left(\boldsymbol{\mu} = \boldsymbol{\delta}_\beta | K, m, J, C, \Delta\right) \geq 1 - \beta$. The stopping boundaries f_j, e_j are then determined by one-dimensional optimisation to find the value of C that minimises $(\alpha - \alpha^*)^2$. With α^* obtained and C fixed, $1 - \beta^*$ is found for some small initial n, which is increased until the required power is reached.

Though we choose to control type-I error-rate and power at one particular point each, $R(\boldsymbol{\mu} | K, m, J, C, \Delta) \leq \alpha$ and $R(\boldsymbol{\mu} | K, m, J, C, \Delta) \geq 1 - \beta$ for (two different) K-dimensional regions. One may be interested in not only controlling type-I error-rate and power at a single point, but across certain regions. This idea is explored further in Section 4.4.1.

With regards to powering the trial for a certain point $\boldsymbol{\delta}_\beta$, we specify anticipated lower and greater effect sizes for each outcome, $\boldsymbol{\delta_0} = (\delta_{01}, \delta_{02}, \ldots, \delta_{0K})$ and $\boldsymbol{\delta_1} = (\delta_{11}, \delta_{12}, \ldots, \delta_{1K})$. We then set $\boldsymbol{\delta}_\beta = (\delta_{11}, \ldots, \delta_{1m}, \delta_{0(m+1)}, \ldots, \delta_{0K})$. That is, exactly m outcomes are equal to their greater anticipated effect δ_{1k}, while $K - m$ outcomes are equal to their lower anticipated effect δ_{0k}. This is analogous to the least favourable configuration (LFC) described by Thall et al. [110] in the context of multi-arm trials. In such trials, the probability of correctly concluding not only that a promising treatment exists, but also identifying that treatment, is of obvious importance. However, in the context of a single-arm trial with multiple outcomes, we place prime importance on the probability of correctly concluding that some subset of m

or more outcomes show promise, rather than additionally correctly identifying the outcomes in this subset. In some situations, it may be of great importance to correctly identify the outcomes that show promise. If so, we can redefine power as the probability of both rejecting the null hypothesis when at least m outcomes have a promising effect size and correctly identifying m of those outcomes.

Above, the m "working" outcomes are taken to be simply the first m outcomes, without loss of generality. They may alternatively be set to be the m smallest standardised outcome effects $\delta_{1k}/\sigma_k, k = 1,\ldots,K$. This may be of use when the anticipated outcome effects, or anticipated variances, differ. In such a case, it would be desirable to power a trial to correctly conclude that m outcomes show promise when such promising outcomes have the m smallest standardised anticipated effects; in single-outcome trials, identifying small effects requires a larger sample size than identifying large effects, and so power is minimised when the m promising outcomes are those with the m smallest standardised effect sizes.

4.3.3 Integration vs. simulation

For both multi-outcome approaches, simulation rather than integration is used to obtain design realisations and their operating characteristics. Grayling et al. [71] present the following notation that fully characterises the progress and conclusion of a MAMS design based on K outcome-specific hypotheses H_k, $1,\ldots K$: $\boldsymbol{\Psi} = (\Psi_1, \Psi_2, \ldots, \Psi_K)$, $\boldsymbol{\Omega} = (\omega_1, \omega_2, \ldots, \omega_K)$, where $\Psi_k = 1$ if H_k is rejected, $\Psi_k = 0$ otherwise and $\Omega_k = j$ where j is either the stage at which H_k is rejected or not rejected or where the trial is stopped. In our multi-outcome multi-stage approach, the test statistics of all outcomes at stage j must considered simultaneously. There are no outcome-specific hypotheses that may be rejected independently of others. It is not sufficient to know that an outcome has crossed a boundary: an outcome may cross a boundary and no trial decision is taken. It is necessary to know the *state* of each outcome's test statistic, that is, which boundary it has crossed (if any), at every stage. As such, it is not

possible in this approach to characterise a trial's progress using two K-length vectors. What is required is J K-length vectors or a $J \times K$ matrix, for example:

$$P = \begin{bmatrix} \Psi_{11} & \Psi_{12} & \dots & \Psi_{1K} \\ \Psi_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \Psi_{J1} & \dots & \dots & \Psi_{JK} \end{bmatrix}, \text{ where } \Psi_{jk} = \begin{cases} 1 & \text{if } Z_{jk} > e_j \\ 0 & \text{if } e_j \geq Z_{jk} \geq f_j \\ -1 & \text{if } Z_{jk} < f_j \end{cases}.$$

In this matrix, each row represents a single stage. As in Grayling et al. [71], the probability of a particular instance of trial progress can be found through a JK-dimensional integration. Each $\Psi_{jk} = \{-1, 0, 1\}$ as defined above has three possible states, and so there are a maximum of 3^{JK} possibilities for the progress of the trial, akin to the "paths" of binary outcome trials described in Chapter 2, and the probability of each can be calculated using the corresponding JK-dimensional integration. The number of possibilities of interest, and so the number of JK integrations required, can be reduced from 3^{JK}. For example, the probability that a trial will end at the first stage need only consider possible states at stage 1. Other reductions are possible, but the degree of reductions required may need to be considerable to manage even a modest trial of of $J = 3$ stages and $K = 3$ outcomes ($3^9 = 19683$ multiple integrals). Conversely, on a computer with an i7-3770 processor and 16GB RAM with no parallelisation, it is possible to simulate 10^5 multi-outcome multi-stage trials of our approach in under 10 seconds.

4.3.4 Design search

We seek to obtain the design realisation that minimises N while satisfying the required type-I error-rate and power. As stated above, the design search for this approach uses simulation. Specifically, we simulate aggregated trial results by simulating JK test statistics, representing the test statistic at each stage j and for each outcome k. We simulate from a multivariate

normal distribution consisting of a mean that is a null vector of length JK and the covariance matrix in Equation (4.6). This approach was suggested by Wason and Jaki [111].

The remaining components of the design search for this approach are described using pseudocode below. Briefly, an optimiser is used in conjunction with Algorithm 4 to find the constant C and corresponding set of lower and upper boundaries that minimise $(\alpha - \alpha^*)^2$. This means that the final design will have a type-I error-rate $\alpha^* \approx \alpha$. To strictly ensure $\alpha^* \leq \alpha$, one might choose to find boundaries by minimising the discontinuous function $(\alpha - \alpha^*)^2$ if $\alpha^* \leq \alpha$, 1 if $\alpha^* > \alpha$. With α^* obtained and C fixed, $1 - \beta^*$ is found for some small initial n, which is increased until the required power is reached. This is shown in Algorithm 5.

4.3.5 Composite outcome design

A simple composite outcome can be created at each stage j by summing the K test statistics Z_{jk}. Let the composite test statistic at stage j be $Z_j = \sum_{k=1}^{K} Z_{jk}$. Each Z_{jk} has been standardised (see Section 4.3), therefore each Z_j is standardised. By taking the sum of the outcomes, all outcomes are being weighted equally. An investigator may choose to apply unequal weights to the outcomes. We undertake a design search analogous to the multi-outcome design search described above, again to find the design realisation that satisfies type-I error-rate and power while minimising N. The same simulated data is used, with the test statistics for each outcome summed to create a composite test statistic for each stage j as described. Again an optimiser is used in conjunction with Algorithm 4 to find some constant C_{COMP} and corresponding stopping boundaries that result in an acceptable type-I error-rate, that is, $\alpha^* \leq \alpha$. The procedure in Algorithm 5 is then used to find the smallest sample size N that will result in an acceptable power, that is, $1 - \beta^* \geq 1 - \beta$.

Algorithm 4: Function pReject: for finding $R()$ and expected number of stages.
Input: $C, J, K, m, \Delta, \alpha, TS$ (matrix of test statistics)

```
lower.bounds ← call findLowerBounds(C, J, Δ);
upper.bounds ← call findUpperBounds(C, J, Δ);
nsims ← nrow(TS) ;
for each i in 1 to nsims do
    for each j in 1 to J do
        TS.current.stage ← call findCurrentTSStage(TS, i, J, K);
        if sum(TS.current.stage > upper.bounds[j]) ≥ m then
            go[i, j] ← 1;
            nogo[i, j] ← 0;
        else
            if sum(TS.current.stage < lower.bounds[j]) ≥ K − m + 1 then
                nogo[i, j] ← 1;
                go[i, j] ← 0;
            else
                go[i, j] ← 0;
                nogo[i, j] ← 0;
            end
        end
    end
    go.nogo.decision[i] ← call findEarliestDecision(go[i, ], nogo[i, ]) ;
    stop.stage[i] ← call findStageOfEarliestDecision(go[i, ], nogo[i, ]);
end
go.decision.count ← sum(go.nogo.decision)==“go” ;
p.reject.null ← go.decision.count/nsims ;
expected.stages.count ← sum(stop.stage)/nsims;
if exists(α) then
    value.to.minimise ← (p.reject.null − α)^2;
end
```

Algorithm 5: Find optimal C for type-I error-rate control, then find smallest value of N that satisfies $1-\beta^* \geq 1-\beta$. Input: $J, K, m, \Delta, \alpha, TS, \boldsymbol{\delta}_\beta, \beta, \boldsymbol{\sigma}$, nmin.

$C \leftarrow$ call optimise(pReject($J, K, m, \Delta, \alpha, TS$));
typeIerror $\leftarrow$ call pReject($C, J, K, m, \Delta, \alpha, TS$);
$\boldsymbol{\mu} \leftarrow \boldsymbol{\delta}_\beta$ `// mu can be set to any vector, if another definition of power is desired;`
pow $\leftarrow 0$;
n.current $\leftarrow$ nmin-1;
while pow $< 1-\beta$ **do**
  n.current $\leftarrow$ n.current+1;
  $\mathscr{I} \leftarrow$ call findInformation(current.n, $\boldsymbol{\sigma}$);
  $\boldsymbol{\tau} \leftarrow$ call findEffects($\boldsymbol{\mu}, \mathscr{I}$);
  TS.current $\leftarrow$ call addEffectsToTS($\boldsymbol{\tau}, TS$);
  pow $\leftarrow$ call pReject(C, n.current, TS.current, $1-\beta$)
end

4.3.6 Comparing multi-outcome and composite designs

The multi-outcome and composite approaches were compared by obtaining design realisations that satisfied the required type-I error-rate and power, set at $\alpha = 0.025$ and $1-\beta = 0.8$. These α and $1-\beta$ were chosen to align with those used in Sozu et al. [49]. The anticipated outcome effect sizes were set as $\delta_{01} = \delta_{02} = \cdots = \delta_{0K} = \delta_0 = 0.2$ and $\delta_{11} = \delta_{12} = \cdots = \delta_{1K} = \delta_1 = 0.4$, again in alignment with Sozu et al., and $\boldsymbol{\delta}_\beta = (\delta_{11}, \ldots, \delta_{1m}, \delta_{0(m+1)}, \ldots, \delta_{0K})$. as described in Section 4.3.2. For simplicity, the variance of each outcome is fixed and equal to one, that is, $\sigma_k^2 = \sigma^2 = 1, \forall k$. $\Delta = 0$ is used in the calculation of stopping boundaries, equivalent to the stopping boundaries proposed by O'Brien and Fleming [107]. The reported operating characteristics are the probability of rejecting the null hypothesis and ESS under the LFC.

We firstly compare rejection regions for single-stage multi-outcome and composite designs. This is followed by comparing design realisations for varying values of correlation ρ. Correlation $\rho_{k_1 k_2} = \rho$, $k_1 \neq k_2$ between all outcomes was equal, and the values examined were $\rho \in \{0, 0.1, \ldots, 0.8\}$.

It is also of interest to examine the consequences of specifying different true outcome effects, given some anticipated outcome effects $\boldsymbol{\delta_\beta}$. When the true effect sizes differ from the effect sizes anticipated in the designs, the performance of both designs will be affected. While we may anticipate which approach may perform better under certain conditions, we wish to quantify these relative changes in performance. We therefore search for design realisations as described above, for both multi-outcome and composite approaches, and note the effect of changing the true outcome effects $\boldsymbol{\mu}$. The required type-I error-rate and power and anticipated outcome effect sizes specified above $(\alpha, \beta, \delta_0, \delta_1, \boldsymbol{\delta}_\beta)$ were also used here, with a shared correlation $\rho_{k_1 k_2} = \rho = 0.3, k_1 \neq k_2$.

4.4 Results: Multi-outcome multi-stage design with general number of required efficacious outcomes

4.4.1 Comparison of single-stage rejection regions

The multi-outcome and composite design approaches lead to different rejection regions. An example of this is shown in Figure 4.2, where a design realisation for each approach has been obtained and the final rejection regions overlaid. The outcome design parameters were $\{K = 2, m = 1, J = 1\}$, that is, single-stage designs. For a composite design, let the lower and upper stopping boundaries for a trial of j stages be $\mathbf{f}^{(c)} = (f_1^{(c)}, f_2^{(c)}, \ldots, f_J^{(c)})$ and $\mathbf{e}^{(c)} = (e_1^{(c)}, e_2^{(c)}, \ldots, e_J^{(c)})$ respectively. For this particular composite design, where $K = 2, J = 1$, the null hypothesis will be rejected at the end of the trial *iff* the sum of the test statistics Z_{11}, Z_{12} is greater than some corresponding efficacy boundary $e_1^{(c)}$, or in general, $\left(\sum_{k=1}^{K} Z_{Jk}\right) > e_J^{(c)}$. For this particular multi-outcome design, the null hypothesis will be rejected at the end of the trial *iff* either test statistic exceeds some corresponding efficacy boundary e_1, or in general, $\left(\sum_{k=1}^{K} \mathbb{I}(Z_{Jk} > e_J)\right) \geq m$. Thus for a general number of outcomes K, rejection of the null hypothesis using the composite design is dependent on all K outcome

test statistics, while rejection of the null hypothesis using the multiple-outcome design occurs if the test statistics of any m outcomes show sufficient response.

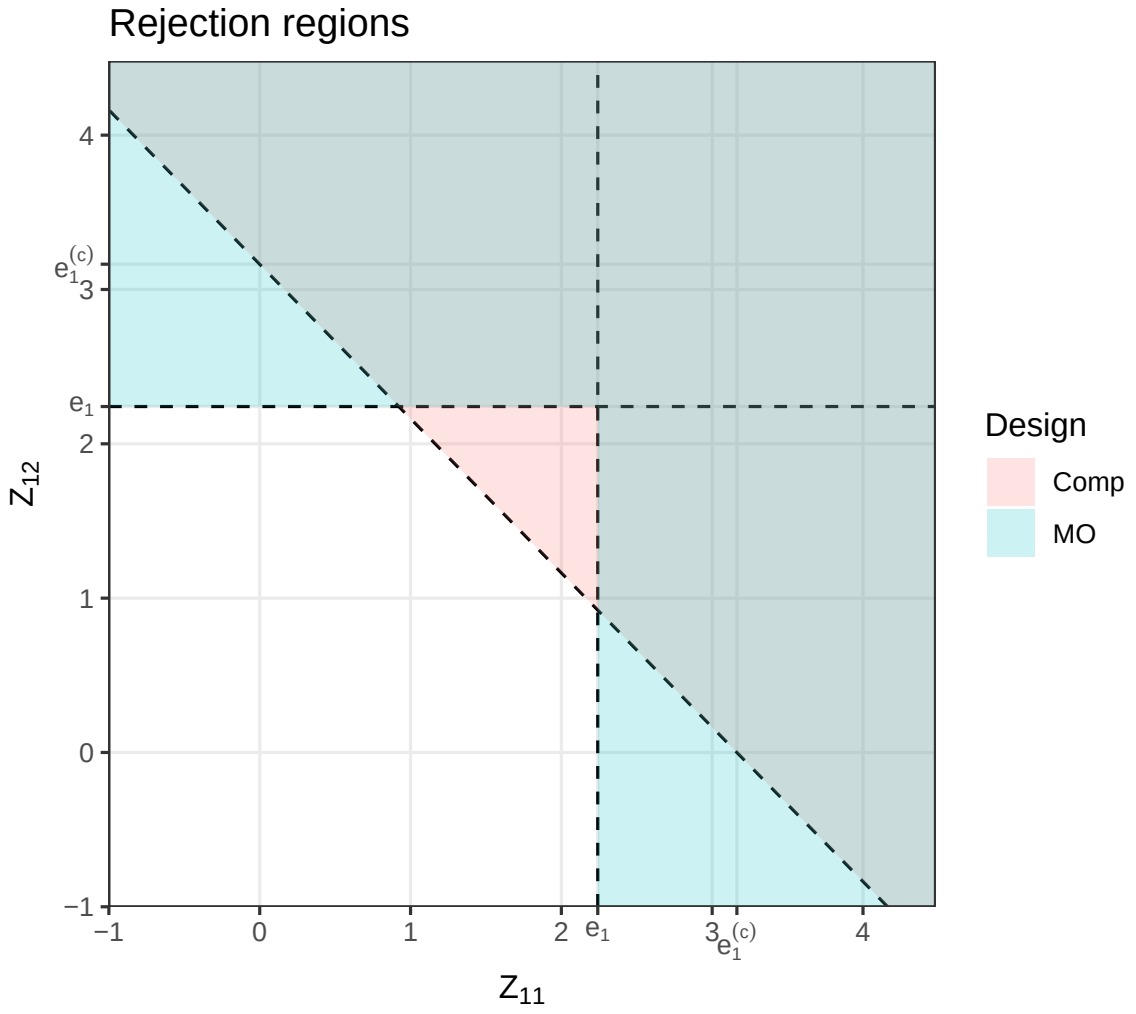


Fig. 4.2 Comparison of final rejection regions for multi-outcome design (blue) and composite design (red), for $\{K = 2, m = 1, J = 1\}$.

The true effect sizes of the outcomes may differ from those specified in the design, and the nature of these differences may affect the performance of the designs in different ways. For example, we expect the multi-outcome approach to outperform the composite approach when some outcomes have a harmful ($\mu_k < 0$) effect, as these outcome effects will dilute any positive effects observed on the remaining outcomes. The opposite effect may occur when more than m outcomes have some moderate effect. In this case, these moderate effects may

combine under the composite design to increase the probability of rejecting null hypothesis compared to the multi-outcome design. We also expect the multi-outcome approach to perform better than the composite approach when fewer than m outcomes have large effect sizes, as the additive aspect of the composite design may cause these outcomes' effects to outweigh the lack of effects in the remaining outcomes, again increasing the probability of rejecting null hypothesis compared to the multi-outcome design. Conversely, we expect the composite approach to perform better when more than m outcomes true effect sizes at least as great as those anticipated, for the same reason. In this case, the outcomes' large effects would make correct rejection of the null hypothesis more likely.

4.4.2 Varying correlation

Figure 4.3 compares the multi-outcome design to the composite design in terms of ESS under the LFC. Define ESS_{MO} and ESS_{comp} as the ESS under the LFC for the multi-outcome and composite designs respectively. The ESS ratio ESS_{MO}/ESS_{comp} under the LFC is shown as correlation ρ varies ($\rho \in \{0, 0.1, \ldots, 0.8\}$). The number of stages was $J = 3$, with the following sets of $\{K, m\}$: $\{K = 2, m = 1\}, \{K = 4, m = 2\}, \{K = 6, m = 1\}, \{K = 6, m = 3\}, \{K = 10, m = 5\}$. A value of less than 1 means that the ESS under LFC is smaller for the multi-outcome design compared to the composite design. Also of interest is the ENM for a given design. In these two approaches, all K outcomes are measured for n participants at each stage j that takes place. As such, ENM in both approaches is simply $K \times ESS$, and so $ESS_{MO}/ESS_{comp} = ENM_{MO}/ENM_{comp}$. ESS ratio decreases as correlation increases. This means that when correlation is low, ESS is relatively poorer on the multi-outcome design, while when correlation is high, ESS is relatively better on the multi-outcome design. The change in ESS ratio as correlation varies is overwhelmingly due to the change in ESS_{comp} as correlation increases. While ESS increases with correlation for both approaches (Figure 4.3, right), the increase is greater for the composite design. For the composite designs found, as

correlation increases, so too does the constant C that determines the stopping boundaries. The boundaries are chosen to ensure the correct type-I error-rate. As the composite test statistic is the sum of the outcome test statistics, increased correlation between outcomes makes type-I errors more likely. Using more extreme upper boundaries counteracts this to ensure an appropriate type-I error-rate. However, using the boundaries of Wang and Tsiatis [105] means that extreme upper boundaries are accompanied by extreme lower boundaries. Two contrasting examples are shown in Figure 4.4, using $\Delta = 0$ as for all design searches in this chapter. The maximum sample size N chosen is the smallest N that results in adequate power. However, with high upper boundaries resulting from having highly correlated outcome test statistics, N must be increased to ensure the design has adequate power.

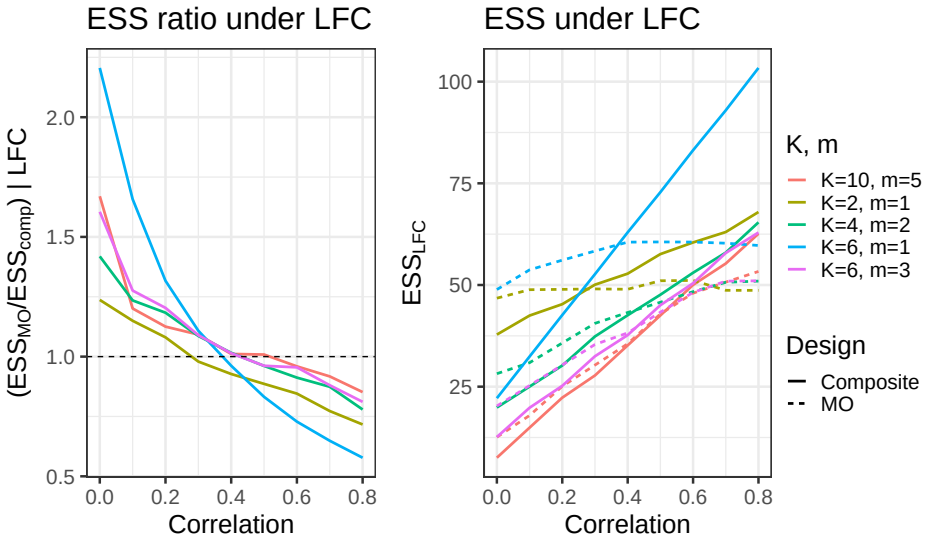
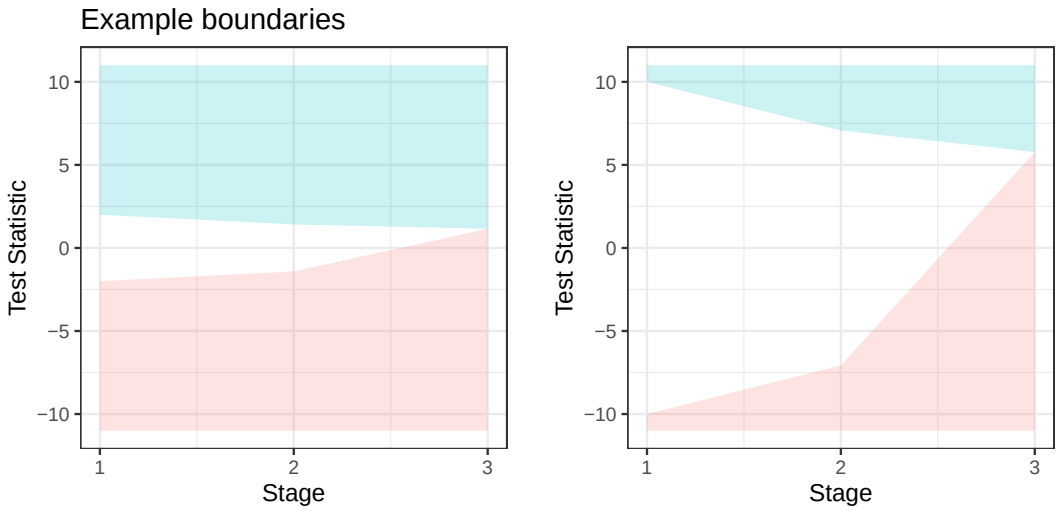


Fig. 4.3 Change in ESS_{MO}/ESS_{comp} as correlation varies. Required error-rates $\alpha = 0.025$, $\beta = 0.2$, design parameters $J = 3, \boldsymbol{\delta}_0 = 0.2, \boldsymbol{\delta}_1 = 0.4$. Simulations: 10^5.

The disparity in ESS between the methods is greatest when $K = 6, m = 1$, the only instance where $m/K < 0.5$. The improvement in ESS under the multi-outcome design compared to the composite design as correlation increases is similar for the remaining combinations, where $m/K = 0.5$. Among these combinations, those with a smaller number of outcomes K appear to benefit more from using a multi-outcome approach compared to



(a) C=2 (b) C=10

Fig. 4.4 Examples of Wang and Tsiatis boundaries for three-stage trials, using Δ=0 and $C = \{2, 10\}$.

a composite approach. For these composite designs, stopping boundaries are independent of m, as type-I error-rate (which is driven by the stopping boundaries) is calculated under the global null. Therefore the boundaries for, say, $\{K = 6, m = 1\}$ and $\{K = 6, m = 3\}$ differ only due to simulation error. However, power is calculated under the LFC, $\boldsymbol{\mu} = \boldsymbol{\delta}_{\beta}$. As such, the observed outcome effects are greater as m increases. When m/K is small, for example, when $\{K = 6, m = 1\}$, rejecting H_0 is less likely, and so N increases to compensate for the lack of power. This explains the larger sample size for the composite design using design parameters $\{K = 6, m = 1\}$. Furthermore, when correlation is high, the $K - m$ null effects are less likely to contribute enough to the composite test statistic to increase power, exacerbating the need for a larger sample size.

4.4.3 Varying true outcome effects

In Figure 4.5, the ESS ratio is compared for a range of different true effects, for a single design realisation of each approach with $\{K = 2, m = 1, J = 3\}$. In this case, the ESS ratio was obtained for every combination of true effect sizes $\mu_1, \mu_2 \in \{-0.2, -0.1, \ldots, 0.4\}$,

with anticipated effect sizes $\boldsymbol{\delta}_\beta = (0.4, 0.2)$ and design parameters $\alpha = 0.025$, $\beta = 0.2$, $\rho_{k_1 k_2} = \rho = 0.3$, $k_1 \neq k_2$, $\sigma_k^2 = \sigma^2 = 1, \forall k$. Using $K = 2$ allows a grid of results to be plotted. Correlation $\rho = 0.3$ is the point in Figure 4.3 at which ESS ratio is close to one for $\{K = 2, m = 1\}$. The design realisations are $\{N = 57, C = 2.256490\}$ for the multi-outcome design and $\{N = 60, C = 3.240066\}$ for the composite design. Across all (μ_1, μ_2) combinations in Figure 4.5, ESS is generally relatively lower when using the multi-outcome design, including the case where the true effect sizes are as anticipated $(\mu_1 = 0.4, \mu_2 = 0.2)$, though at this point the ESS ratio is close to one. The only regions where ESS is greater using the multi-outcome design is when the "non-working" outcome has a greater than anticipated effect size $(\mu_2 > 0.2)$ and when both outcomes are particularly harmful $(\mu_1 = -0.2, \mu_2 = -0.2)$. In the former case, the composite design is more likely to reject H_0 sooner as the design combines the positive observed effects of both outcomes. Similarly, in the latter case, the two negative observed effects combine, resulting in a test statistic that causes a trial to end for a no go decision sooner than the corresponding multi-outcome design.

Figure 4.6 shows the how the probability of rejecting H_0 changes for different true effects, for the same multi-outcome and composite design realisations as Figure 4.5. When using the multi-outcome design, P(reject H_0) remains at least close to the required power when either outcome has true effect $\mu = 0.4$, while when using the composite design, P(reject H_0) decreases below the required power when one outcome has true effect $\mu = 0.4$ and the other has some true effect less than 0.2. As above, the combining of outcome effects on the composite design is responsible for this, with the lower-than-anticipated observed effect "cancelling out" the positive observed effect to some extent. This can be seen in Figure 4.2, where a low value for test statistic Z_{11} (or Z_{12}) means that a greater test statistic Z_{12} (or Z_{11}) is required to reject H_0 under the composite design but not the multi-outcome design.

Table 4.1 also compares the two approaches in terms of a single design realisation for each approach, for a range of different true effects. In this case, the total number of

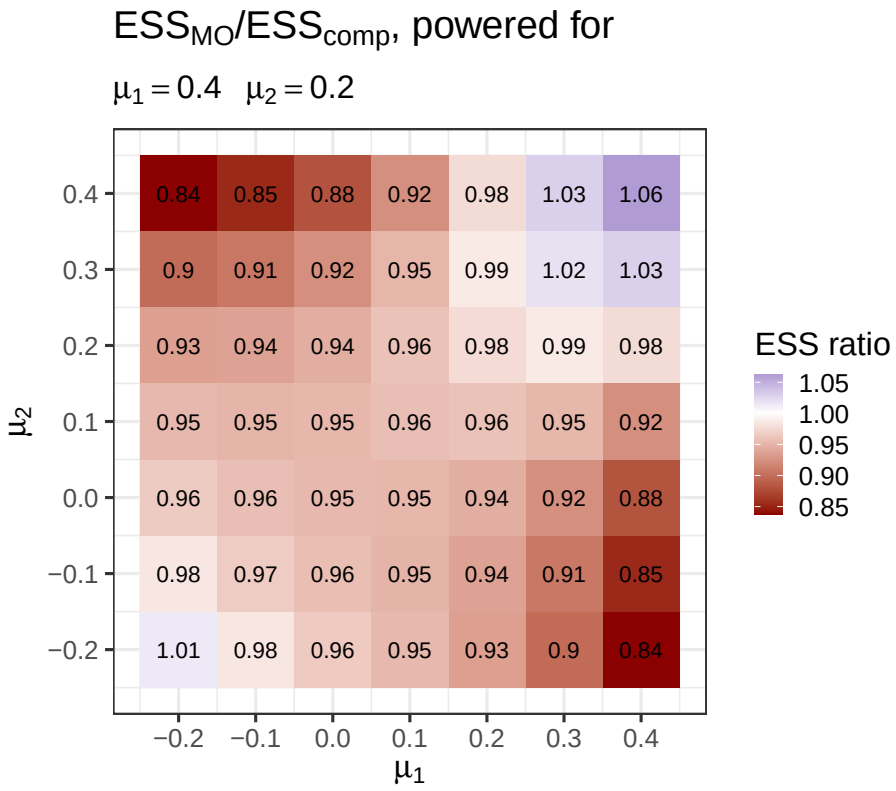


Fig. 4.5 Change in ESS_{MO}/ESS_{comp} as true outcome effect sizes vary. Required error-rates $\alpha = 0.025$, $\beta = 0.2$, design parameters $\{K = 2, m = 1, J = 3\}, \delta_{\beta 1} = 0.4, \delta_{\beta 2} = 0.2, \rho = 0.3$. Simulations: 10^5.

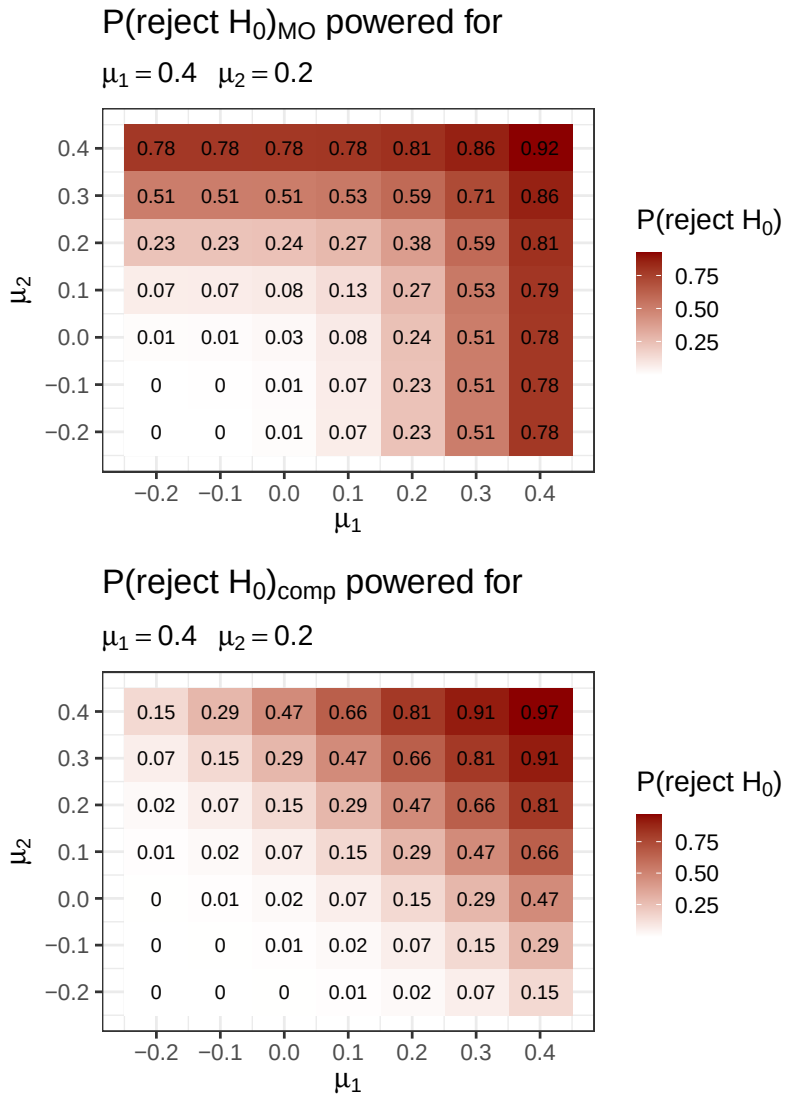


Fig. 4.6 $R(\boldsymbol{\mu} = \mu_1, \mu_2)$ as true outcome effect sizes vary. Required error-rates $\alpha = 0.025$, $\beta = 0.2$, design parameters $\{K = 2, m = 1, J = 3\}$, $\boldsymbol{\delta}_\beta = (0.4, 0.2)$. Simulations: 10^5.

outcomes is increased to $K = 3$ while the remaining design parameters are unchanged. The design realisations are $\{N = 60, C = 2.394350\}$ for the multi-outcome design and $\{N = 63, C = 4.387731\}$ for the composite design. The ESS ratio is examined again, as is the probability of rejecting the null hypothesis, for a range of scenarios. The ESS ratio is greater than or equal to 1, i.e. ESS is poorer under the multi-outcome design, when all outcomes have equal non-zero true effects. Here, the composite design benefits from combining the observed effects. The ESS ratio is less than 1 otherwise, favouring the multi-outcome design. The relative difference in favour of the multi-outcome design is at its greatest when the "non-working" outcomes have a zero or harmful true effect, where the composite design either does not benefit or is even harmed by combining outcome effects. As in Figure 4.6, under the multi-outcome design P(reject H_0) is close to the nominal power (or greater) when at least one outcome has a true effect equal to the anticipated effect, while under the composite design P(reject H_0) decreases as the true effect sizes of the "non-working" outcomes decrease, even if one outcome has a true effect equal to the anticipated effect. When all three outcomes have some true effect that is lower than δ_{1k}, e.g. $\mu_1 = \mu_2 = \mu_3 = 0.3$ or $\mu_1 = \mu_2 = \mu_3 = 0.2$, rejecting the null hypothesis is more likely under the composite design than the multi-outcome design. Again, the multi-outcome design will only reject the null upon observing effects of a particular size on m outcomes only, while the composite design may reach the rejection region by combining these smaller observed effects.

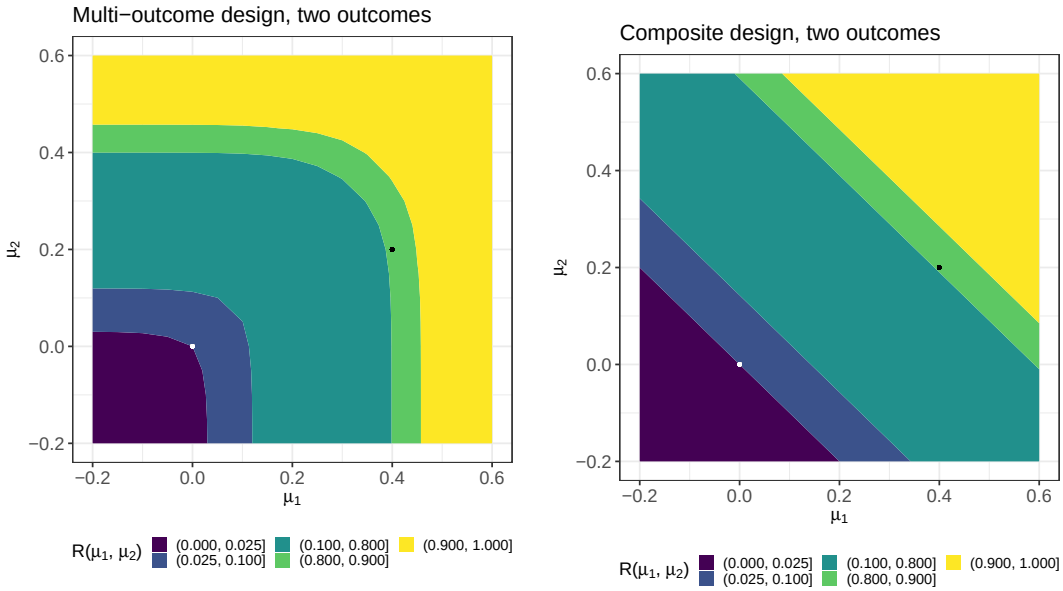
μ_1	μ_2	μ_3	$R(\boldsymbol{\mu})_{MO}$	$R(\boldsymbol{\mu})_{comp}$	$ESS_{MO/comp}$	Description
0.4	0.4	0.4	0.96	0.99	1.13	All outcomes have effect δ_1
0.4	0.2	0.2	0.81	0.82	0.99	Effects as anticipated (power)
0.4	0.0	0.0	0.76	0.30	0.87	Two outcomes have no effect
0.4	-0.2	-0.2	0.76	0.02	0.84	Two outcomes are harmful
0.0	0.0	0.0	0.02	0.02	0.96	Global null (type-I error)
0.3	0.3	0.3	0.78	0.90	1.07	All have some effect $< \delta_1$
0.2	0.2	0.2	0.44	0.58	1.00	All outcomes have effect δ_0

Table 4.1 $R(\boldsymbol{\mu} = \mu_1, \mu_2, \mu_3)$ and expected sample size ratios for MO design and composite design, where $K = 3, m = 1, J = 3, \boldsymbol{\delta}_\beta = (0.4, 0.2, 0.2)$.

The idea that multi-outcome designs have rejection regions or spaces was introduced in Section 4.3.2. We compare the different rejection regions of the multi-outcome multi-stage design with the composite design for $\{K = 2, m = 1, J = 3\}$ and $\{K = 3, m = 2, J = 3\}$. In Figure 4.7, we show $R(\mu_1, \mu_2)$, the probability of rejecting H_0 given true outcome effects μ_1, μ_2. The required error-rates were $\alpha = 0.025, \beta = 0.2$, with the designs powered for outcome effect sizes $\mu_1 = 0.4, \mu_2 = 0.2$. Suitable design realisations were obtained for $N = 57$ (19 per stage) in the multi-outcome design and $N = 60$ (20 per stage) for the composite design (as above). For this comparison, we wanted N to be equal for the design realisations of both designs. Requiring $N = 60$ for the multi-outcome design meant that power was increased, hence the power is greater than may be expected ($1 - \beta^* = 0.827$, black dot on Figure 4.7a). The black dots, representing the points for which the designs are powered, are do not lie exactly on a contour. Beyond the explanation for the increased power of the multi-outcome design above, this is due to the discrete nature of sample size: for these designs, sample size is increased until the required power is reached. The type-I error-rate is determined by the stopping boundary constant C, which may take any continuous value. Consequently, the white dots, indicating the point at which the type-I error-rate must be satisfied, both lie exactly on a contour.

The shapes of the regions largely reflect those in Figures 4.2 and 4.6: the group of regions within which the probability of rejection is low is approximately square for the multi-outcome design and triangular for the composite design. The reasoning remains the same: the multi-outcome design does not penalise a negative effect size, unlike the composite design. In general, the additive nature of the composite design plays a strong role in the differences between the regions.

Figure 4.8 shows rejection regions for three outcomes, powered to find two promising outcomes $\boldsymbol{\delta}_\beta = (0.4, 0.4, 0.2)$ and with three stages, that is, $\{K = 3, m = 2, J = 3\}$. The sample size on the composite design was increased so that sample size was equal across

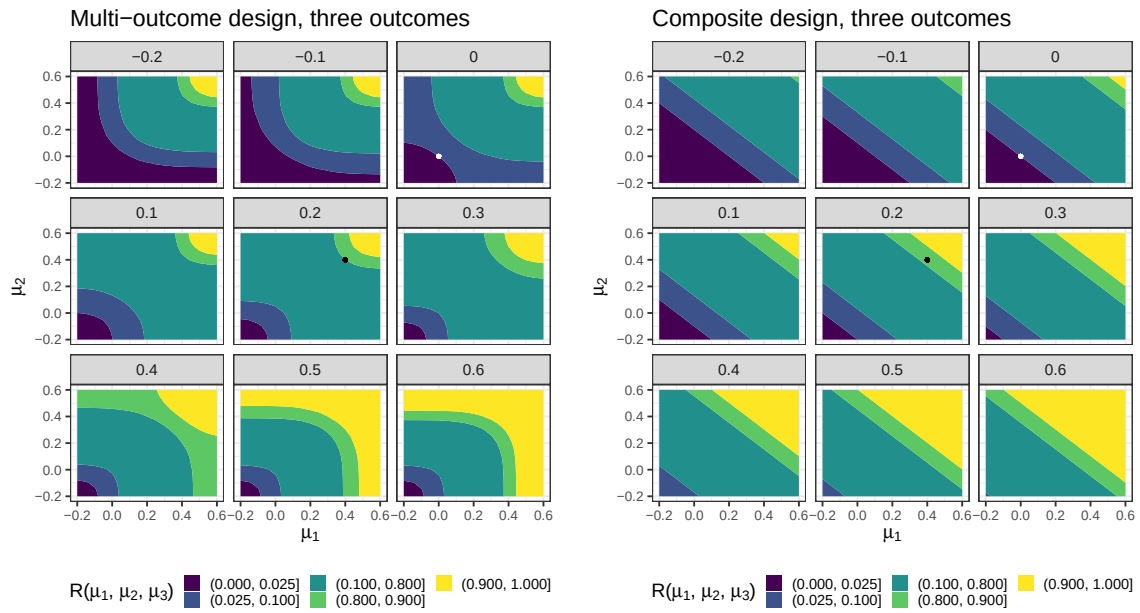


(a) Multi-outcome multi-stage design realisation with $N = 60, C = 2.256490, \Delta = 0$. Operating characteristics: $\alpha^* = 0.025, 1 - \beta^* = 0.827$.

(b) Composite multi-stage design realisation with $N = 60, C = 3.240066, \Delta = 0$. Operating characteristics: $\alpha^* = 0.025, 1 - \beta^* = 0.814$.

Fig. 4.7 Probability of rejecting H_0 as true effect sizes vary. Powered for effect sizes $\boldsymbol{\delta}_\beta = (0.4, 0.2)$. Design parameters $K = 2, m = 1, J = 3, \alpha = 0.025$, $\beta = 0.2$, $\rho_{k_1 k_2} = \rho = 0.3, k_1 \neq k_2$, $\sigma_k^2 = \sigma^2 = 1, \forall k$. White dot indicates global null $\boldsymbol{\mu} = \mathbf{0}$, black dot indicates point for which design is powered, $\boldsymbol{\mu} = \boldsymbol{\delta}_\beta$.

both design realisations, $N = 42$ (14 per stage). Again, the black dots indicating power do not lie exactly on a contour, in contrast with the white dots indicating type-I error-rate which do lie exactly on a contour. For $\mu_3 \in \{0.5, 0.6\}$, rejection regions are similar to Figure 4.7 (and again Figures 4.2 and 4.6). For non-positive values of μ_3, the low probability of rejecting H_0 in the composite design can be seen, even when the remaining outcomes have considerable effect sizes. Conversely, for the corresponding plots on for multi-outcome design, non-positive values of μ_3 have little effect on the size of the rejection regions. This again shows the nature of the difference between an additive and non-additive test statistic. As μ_3 increases, the rejection regions of the composite design seem to shift linearly and without changing shape. However, in the multi-outcome design the regions corresponding to high probability of rejection change shape as μ_3 increases, from a small square to a large inverted "L" shape. Conversely, the region corresponding to low probability of rejection changes shape in the opposite way. This is because when μ_3 is low, there is little chance of this outcome contributing to a rejection of H_0. As μ_3 increases closer to the value for which the promising outcomes are powered, this probability increases. When μ_3 is much greater than this, it is almost certain to contribute to the rejection of H_0 (by exceeding its stopping boundary). As such, only one of the two remaining outcomes μ_1, μ_2 are additionally required to show promise for H_0 to be rejected. Therefore H_0 is likely to be rejected when either one of μ_1, μ_2 shows promise. Furthermore, as rejection of H_0 is dependent on only (any) two outcomes showing an effect, there is little "benefit" from all three outcomes having large effect sizes. Indeed, for this design realisation, the probability of rejecting H_0 when any $\mu_k = \infty, \mu_j = 0$ for $k \in \{1,2,3\}, j \neq k$, is approximately 0.12, while this probability is necessarily equal to one for any composite design realisation.



(a) Multi-outcome multi-stage design realisation with $N = 42, C = 1.579395, \Delta = 0$. Operating characteristics: $\alpha^* = 0.025, 1 - \beta^* = 0.801$.

(b) Composite multi-stage design realisation with $N = 42, C = 4.389363, \Delta = 0$. Operating characteristics: $\alpha^* = 0.025, 1 - \beta^* = 0.836$.

Fig. 4.8 Probability of rejecting H_0 as true effect sizes vary. Powered for effect sizes $\boldsymbol{\delta}_\beta = (0.4, 0.4, 0.2)$. Design parameters $K = 3, m = 2, J = 3, \alpha = 0.025$, $\beta = 0.2$, $\rho_{k_1 k_2} = \rho = 0.3, k_1 \neq k_2$, $\sigma_k^2 = \sigma^2 = 1, \forall k$. White dot indicates global null $\boldsymbol{\mu} = \mathbf{0}$, black dot indicates point for which design is powered, $\boldsymbol{\mu} = \boldsymbol{\delta}_\beta$. Each plot slice represents true effect size for μ_3.

4.5 Methods: Drop the loser approach based on conditional probability, two-stage

The multi-outcome multi-stage approach may be combined with a DtL-type component, that is, dropping an outcome (or outcomes) before the end of trial, with the aim of reducing ENM. This approach to reducing the number of measurements in a trial is an alternative to using separate stopping rules, as described in Section 4.3.

Again, let K be the number of outcomes and m be the number of outcomes required to show promise in order to reject the null hypothesis. Fix the number of stages to be equal to 2. The shared final rejection boundary is given by r. If an outcome k is not dropped at the interim analysis, that is, it is still being measured at the end of the trial, its test statistic Z_{2k} will be compared against this final rejection boundary r. We again specify lower and greater anticipated treatment effects for each outcome, $\boldsymbol{\delta_0} = (\delta_{01}, \delta_{02}, \ldots, \delta_{0K})$ and $\boldsymbol{\delta_1} = (\delta_{11}, \delta_{12}, \ldots, \delta_{1K})$. Let the true outcome effects again be $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$.

The number of outcomes dropped at the interim analysis may be fixed in advance or determined by the interim data. In either case, some approach must be used to determine the "losers", the poorest-performing outcomes. The approach we have chosen is to use conditional power (CP) [52, 10]. Here, we define CP_k as the probability of outcome k exceeding the final rejection boundary r, conditional on the data for outcome k observed so far and an anticipated outcome effect δ_{1k}. For a general number of stages $j, j = 1, \ldots, J$, the conditional power of outcome k at stage j is then $CP_{jk}(\delta_{1k}) = \mathrm{P}(Z_{Jk} > r | Z_{jk}, \delta_{1k})$. The calculation for the conditional power of outcome k at the single interim analysis, given current data and anticipated outcome effect δ_{1k} is

$$CP_k(\delta_{1k}) = \Phi\left(\frac{Z_{1k}\sqrt{\mathscr{I}_1} - r\sqrt{\mathscr{I}_2} + (\mathscr{I}_2 - \mathscr{I}_1)\delta_{1k}}{\sqrt{(\mathscr{I}_2 - \mathscr{I}_1)}}\right). \tag{4.7}$$

Equation (4.7) is merely a special case of the equation for a general number of stages j provided by Jennison and Turnbull [10]. As in Section 4.3, using a single shared boundary avoids a K-dimensional optimisation problem with infinite solutions.

4.5.1 Conditional power-based stopping (and dropping) boundaries

CP is used in this multi-outcome multi-stage design, rather than comparing test statistics to boundaries directly as may be expected in the multi-arm setting. However, multi-arm trials are generally used to evaluate if any treatment has some single effect size of interest. In contrast, outcomes may have different anticipated effect sizes. As such, absolute values of test statistics may not give an accurate indication of the relative interim performance of outcomes. For example, among two interim test statistics, one test statistic (Z_{11} say) may be lower than another (Z_{12}) while being closer to its anticipated standardised effect size, i.e. $(\delta_{11}/\sigma_1) - Z_{11} < (\delta_{12}/\sigma_2) - Z_{12}, \quad (\delta_{11}/\sigma_1) < Z_{11}, (\delta_{12}/\sigma_2) < Z_{12}$. In this case, the outcome with the lower test statistic may be the outcome that is more likely to exceed the final rejection boundary, and should not necessarily be the outcome that is dropped.

We specify lower and upper interim stopping boundaries in terms of some conditional probabilities CP_L and CP_U. Our approach to dropping outcomes and to stopping the trial are as follows: if the CP of the test statistic of some outcome k is less than CP_L at the interim, that is, $CP_k(\delta_{1k}) < CP_L$, it is dropped from the trial and not measured nor evaluated at the final stage. If $K - m + 1$ or more outcomes are dropped, the trial ends early for a no-go decision. If the CPs of the test statistics of m or more outcomes are greater than CP_U at the interim, that is, if $\sum_{k=1}^{K} \mathbb{I}(CP_k(\delta_{1k}) > CP_U) \geq m$, the trial ends early for a go decision. If the trial does not end early, it proceeds to a second stage. The number of outcomes retained for stage 2 is

$$K_2 = min\left(K_{max}, \sum_{k=1}^{K} \mathbb{I}(CP_k(\delta_{1k}) > CP_L)\right)$$

for some fixed $K_{max} < K$. The value of K_{max} determines the maximum possible number of outcomes that may be measured in stage 2, and thus also determines the maximum number of outcome measurements obtained in this design approach.

The null hypothesis is unchanged compared to the previous multi-outcome multi-stage approach, given by Equation (4.1). The null hypothesis is rejected if either the trial continues to stage 2 and at least m retained outcomes exceed the final stopping boundary r, or if at least m CP values exceed CP_U at the interim, that is if

$$\left(\sum_{k=1}^{K} CP_k(\delta_{1k}) > CP_L \cap \mathbb{I}(Z_{2k} > r)\right) \geq m \qquad \text{and} \qquad \sum_{k=1}^{K} \mathbb{I}(CP_k(\delta_{1k}) > CP_U) < m$$

$$\text{or} \qquad \sum_{k=1}^{K} \mathbb{I}(CP_k(\delta_{1k}) > CP_U) \geq m.$$

As with the first proposed multi-outcome approach, we define the probability of rejecting the null hypothesis for outcome effects $\boldsymbol{\mu}$, but for this approach the design parameters are K, K_{max}, m, CP_L, CP_U. We define type-I error-rate as the probability of rejecting the null hypothesis under the global null, $\alpha^* = R(\boldsymbol{\mu} = \mathbf{0}|K, K_{max}, m, CP_L, CP_U)$ and the power as the probability of rejecting the null hypothesis under the LFC, $1-\beta^* = R(\boldsymbol{\mu} = \boldsymbol{\delta}_\beta|K, K_{max}, m, CP_L, CP_U)$ similar to Section 4.3.

4.5.2 Design search

To search for designs, sets of $2K$ test statistics are simulated under the global null hypothesis $\boldsymbol{\mu} = \mathbf{0}$. A search is undertaken to find a design that fulfils the required type-I error-rate α and power $1-\beta$. The interim boundaries CP_L, CP_U and anticipated effects $\boldsymbol{\delta_0}, \boldsymbol{\delta_1}$ are fixed and specified in advance. The operating characteristics of a trial therefore depend on the final rejection boundary r and the per-stage sample size n. A shared rejection boundary r is found that minimises $(\alpha - \alpha^*)^2$ for some initial per-stage n. Using these boundaries and n, $1-\beta^*$

is obtained. If $1-\beta^*$ is less than the required power $1-\beta$, then the process of finding r and power is repeated with increased n. Conversely if $1-\beta^*$ is greater than the required power $1-\beta$, the process is repeated with decreased n. Thus the per-stage sample size n is altered to find the smallest value that satisfies the required power. Some *nsims* number of trials are simulated as in Section 4.3.4. The rest of the design search is described in Algorithms 6, 7 and 8.

Algorithm 6: findCPs: find conditional power at stage 1, for a vector of outcomes.
Input: TSrow (one row of K simulated interim test statistics), $\mathscr{I}_1, \mathscr{I}_2, \boldsymbol{\delta}_1, r$.

numerator $\leftarrow$ TSrow$\sqrt{\mathscr{I}_1} - r\sqrt{\mathscr{I}_2} + (\mathscr{I}_2 - \mathscr{I}_1)\boldsymbol{\delta}_1$;
denominator $\leftarrow \sqrt{(\mathscr{I}_2 - \mathscr{I}_1)}$;
cp $\leftarrow$ call normalCDF(numerator/denominator);

4.6 Results: drop the loser approach based on conditional probability, 2-stage

4.6.1 Varying correlation

The multiple outcome DtL approach was compared to a multiple outcome single-stage approach in terms of ESS and ENM ratios (denoted ESS_{DtL}/ESS_{single} and ENM_{DtL}/ENM_{single}) under the LFC as correlation varied ($\rho = \{0, 0.1, \ldots, 0.8\}$). Design realisations were found for $\{K, m\} = \{2,1\}, \{6,1\}, \{6,3\}$ and $K_{max} = \{K-1, K/2\}$ (see Table 4.2). Other design parameters were as the previous approach (Section 4.3.6): $\alpha = 0.025, \beta = 0.2, \delta_{01} = \delta_{02} = \cdots = \delta_{0K} = \delta_0 = 0.2, \delta_{11} = \delta_{12} = \cdots = \delta_{1K} = \delta_1 = 0.4, \sigma_k^2 = 1, \forall k$. The lower and upper conditional power thresholds were set to $CP_L = 0.3$ and $CP_U = 0.95$ respectively. In Chapters 2 and 3, the maximum lower threshold for CP was set equal to the response rate for which the trial was powered. Here, there is no such obvious association to be made between CP threshold, a probability and effect size, a continuous value. In the absence of sugges-

Algorithm 7: Function pRejectDTL: for finding type-I error-rate or power, expected number of stages and ENM for DtL approach. Input: r, K, K_{max} n.per.stage (current n per stage), $m, \boldsymbol{\delta}_0, \boldsymbol{\delta}_1, CP_L, CP_U, \boldsymbol{\sigma}^2, \alpha, TS$ (matrix of test statistics), type1err.or.power (whether finding type-I error-rate or power)

$\mathscr{I}_1 \leftarrow$ call findInformation(n.per.stage, $\boldsymbol{\sigma}^2, K$);
$\mathscr{I}_2 \leftarrow$ call findInformation(2*n.per.stage, $\boldsymbol{\sigma}^2, K$);
if type1.err.or.power=="power" **then**
  $\boldsymbol{\delta}_\beta \leftarrow$ call findDeltaBeta($\boldsymbol{\delta}_0, \boldsymbol{\delta}_1$);
  $\boldsymbol{\tau} \leftarrow$ call findEffects($\boldsymbol{\delta}_\beta, \mathscr{I}_1, \mathscr{I}_2$);
  TS $\leftarrow$ call addEffectsToTS(TS, $\boldsymbol{\tau}$);
end
TS.stage1 $\leftarrow$ TS[, 1:K];
TS.stage2 $\leftarrow$ TS[, $(K+1)$:$2K$];
nsims $\leftarrow$ nrow(TS);
for i in 1 to nsims **do**
  CPs[i,] $\leftarrow$ call findCPs(TS.stage1[i,], $\mathscr{I}_1, \mathscr{I}_2, r, \boldsymbol{\delta}_1$);
  if sum(CPs[i,] $< CP_L$) $\geq K - m + 1$ **then**
    no.go.decision.stage1[i] $\leftarrow$ 1 ;
    go.decision.stage1[i] $\leftarrow$ 0 ;
  else
    if sum(CPs[i,] $> CP_U$) $\geq m$ **then**
      no.go.decision.stage1[i] $\leftarrow$ 0 ;
      go.decision.stage1[i] $\leftarrow$ 1 ;
    else
      no.go.decision.stage1[i] $\leftarrow$ 0 ;
      go.decision.stage1[i] $\leftarrow$ 0 ;
    end
  end
end
stop.early $\leftarrow$ no.go.decision.stage1 + go.decision.stage1 ;
continue $\leftarrow$!stop.early ;
TS.continue $\leftarrow$ call subsetToContinuingTrials(TS.stage2, continue);
CPs.continue $\leftarrow$ call subsetToContinuingTrials(CPs, continue);
nrows.continue $\leftarrow$ sum(continue);
for i in 1 to nrows.continue **do**
  CPs.ranked $\leftarrow$ call rankCPs(CPs.continue[i,]);
  retained.outcomes[i] $\leftarrow$ call retainGreatestCPs(CPs.ranked, K_{max}, CP_L);
  retained.TSs $\leftarrow$ call subsetToRetainedTSs(TS.continue[i,], retained.outcomes[i]);
  if sum(retained.TSs $> r$) $\geq m$ **then**
    go.decision.stage2[i] $\leftarrow$ 1 ;
  else
    go.decision.stage2[i] $\leftarrow$ 0 ;
  end
end
no.measurements.stage2 $\leftarrow$ sum(retained.outcomes) ;
prob.reject $\leftarrow$ go.decision.stage1 + go.decision.stage2 ;
PET $\leftarrow$ sum(stop.early)/nsims ;
ENM $\leftarrow$ K + no.measurements.stage2/nsims ;
if type1.err.or.power=="typeIerror" **then**
  minimise.prob $\leftarrow$ (prob.reject $- \alpha)^2$;
end

Algorithm 8: findDtLDesign: find DtL design realisation that satisfies required type-I error-rate and power. Input: TS, nmin, nmax, m, K, K_{max}, $\alpha, 1-\beta$, $\mathscr{I}_1, \mathscr{I}_2, \boldsymbol{\delta}_0, \boldsymbol{\delta}_1, CP_L, CP_U, \boldsymbol{\sigma^2}$.

```
n.vec ← nmin:nmax ;
a ← 1;
b ← length(n.vec);
d ← ceiling((b-a)/2);
while b − a > 1 do
    r.current ← call optimise(pRejectDTL, typeIerror.or.power="typeIerror",
      n.per.stage=n.vec[d], ...);
    pow ← call pRejectDTL(r.current, typeIerror.or.power="power",
      n.per.stage=n.vec[d], ...);
    if pow < power then
        a ← d;
        d ← ceiling(a + (b − a)/2);
    else
        b ← d;
        d ← ceiling(a + (b − a)/2);
    end
end
n.final ← n.vec[d];
r.final ← call optimise(pRejectDTL(n.per.stage=n.final,
  typeIerror.or.power="typeIerror"));
typeIerr.output ← call pRejectDTL(r.final, typeIerror.or.power="typeIerror",
  n.per.stage=n.final, ...);
power.output ← call pRejectDTL(r.final, typeIerror.or.power="power",
  n.per.stage=n.final, ...);
α* ← call selectPReject(typeIerr.output);
pow ← call selectPReject(power.output);
N ← 2*n.final;
ESS0 ← call selectPET(typeIerr.output)*n.final + (1-selectPET(typeIerr.output))*N;
ESS1 ← call selectPET(power.output)*n.final + (1-selectPET(power.output))*N;
ENM0 ← call selectENM(typeIerr.output)*n.final;
ENM1 ← call selectENM(power.output)*n.final;
```

tions in the literature, the thresholds $CP_L = 0.3, CP_U = 0.95$ were chosen. The admissible single-stage designs of Chapter 2 often have similar thresholds, though we acknowledge the difference in the design approaches. At the trial planning stage, we recommend undertaking a sensitivity analysis of the interim thresholds, to more fully understand how the choice may affect a particular set of design parameters. Note that setting $CP_L = 0$ is equivalent to permitting early stopping for a go decision only, while setting $CP_U = 1$ is equivalent to permitting early stopping for a no go decision only.

K	m	K_{max}
2	1	1 ($K_{max} = K-1, K/2$)
6	1	3 ($K_{max} = K/2$)
6	1	5 ($K_{max} = K-1$)
6	3	3 ($K_{max} = K/2$)
6	3	5 ($K_{max} = K-1$)

Table 4.2 Sets of design parameters $\{K, m, K_{max}\}$ used in comparison of proposed DtL design and single-stage design.

The results are shown in Figure 4.9. Values below 1 indicate superiority of the proposed DtL approach over the single stage approach. Similarly to the previous results, ESS ratio decreases as correlation ρ increases. However, in this comparison, the ESS ratio is less than 1 in almost all cases, and in all but one case when $\rho > 0$, though the ESS ratio is generally closer to 1 compared to the results in Figure 4.3. ESS ratio appears to be greater when $m > 1$, though this difference seems to decrease as ρ increases. The ENM ratio also decreases as ρ increases. The ENM ratio is less than 1 in every case, meaning that fewer measurements are expected over both stages of the DtL design than in the single stage design. The ENM ratio is considerably greater under $\{K = 2, m = 1\}$ compared to the other combinations of $\{K, m\}$ examined. Using $K_{max} = K/2$ resulted in a lower ENM ratio than using $K_{max} = K - 1$. This may be expected, as fewer outcomes are permitted to be retained for the second stage.

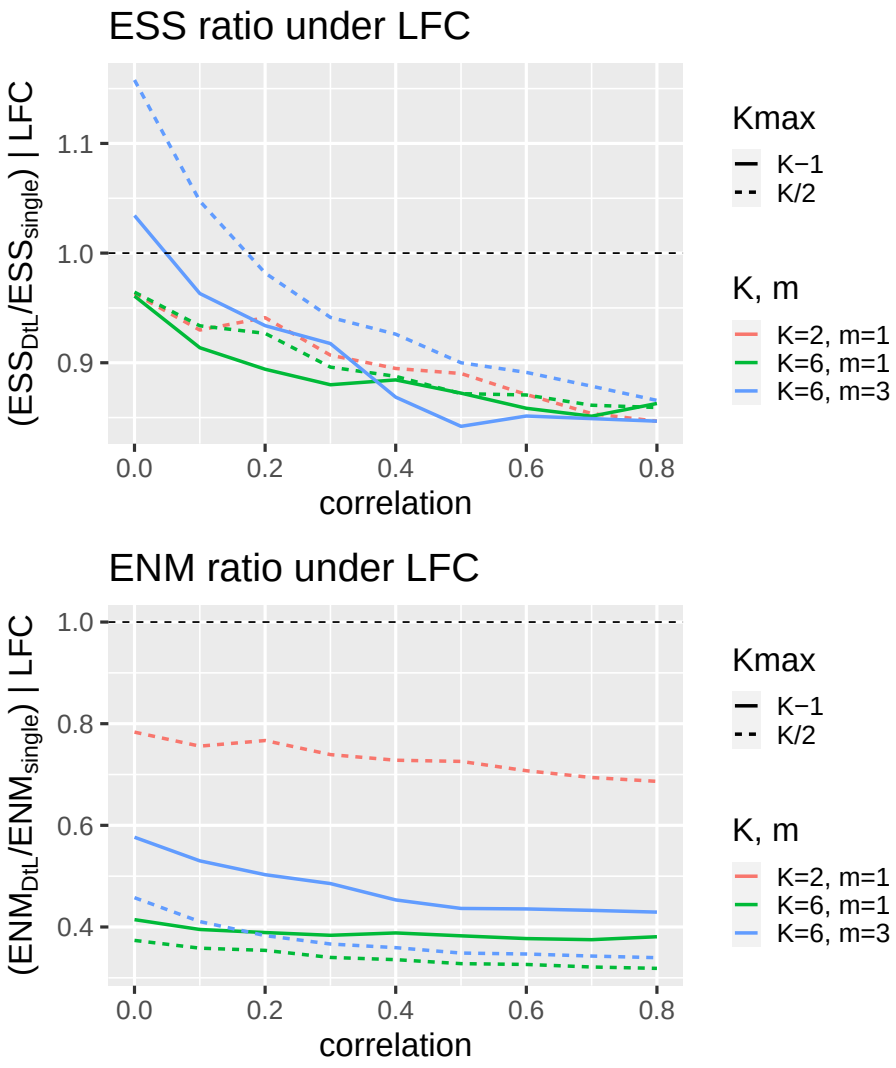


Fig. 4.9 Changes in ESS_{DtL}/ESS_{single} and ENM_{DtL}/ENM_{single} for various designs as correlation ρ is varied. Note: for $\{K = 2, m = 1\}, K - 1 = K/2$.

4.6.2 Varying true outcome effects

4.6.2.1 Two outcomes

The changes in ESS and ENM ratio for single design realisations as true outcome effects vary over $\mu_1, \mu_2 \in \{-0.2, -0.1, \ldots, 0.4\}$ are shown in Figure 4.10. Design parameters are $\{K = 2, K_{max} = 1, m = 1\}, \boldsymbol{\delta}_\beta = (0.4, 0.2)$. The design realisations are $\{r = 2.273714, N = 64\}$ for the DtL design and $\{r = 2.221584, N = 56\}$ for the single-stage design.

As in Figure 4.9, ESS ratio is generally less than 1, with ESS being lower in the single stage design in just three out of 49 cases. This occurs when both $\mu_1, \mu_2 \approx \delta_{\beta 2} = 0.2$. The greatest disparity in ESS is when $\mu_1 = \mu_2 = -0.2$, the minimum effect size examined. When the true effect sizes are low, or even harmful, the conditional power will be low and the possibility of early stopping increases. The ENM ratio shows similar results, with the lowest values (and greatest benefit of the DtL design) observed when the true outcome effects are at their lowest with either trial ending or dropping an outcome at the interim. The ENM ratio is less than 1 in all cases.

For the same design realisations, the probability of rejecting H_0 under each approach is shown for $\mu_1, \mu_2 \in \{-0.2, -0.1, \ldots, 0.4\}$ in Figure 4.11. Both approaches show increases as one or both effect sizes increase. For all cases such that one outcome has the anticipated effect size 0.4 while the other has an effect size of 0.1 or lower, the DtL design reports a probability of rejecting H_0 slightly greater than nominal $[0.80, 0.82]$, possibly due to a slightly increased probability of dropping the poorly-performing outcome over the better-performing outcome compared to having effect sizes of $\boldsymbol{\mu} = (0.4, 0.2)$. For the same cases, the single stage design reports a probability slightly lower than nominal $[0.78, 0.79]$, possibly due to a slightly decreased probability of of rejecting H_0 due to the poorly-performing outcome.

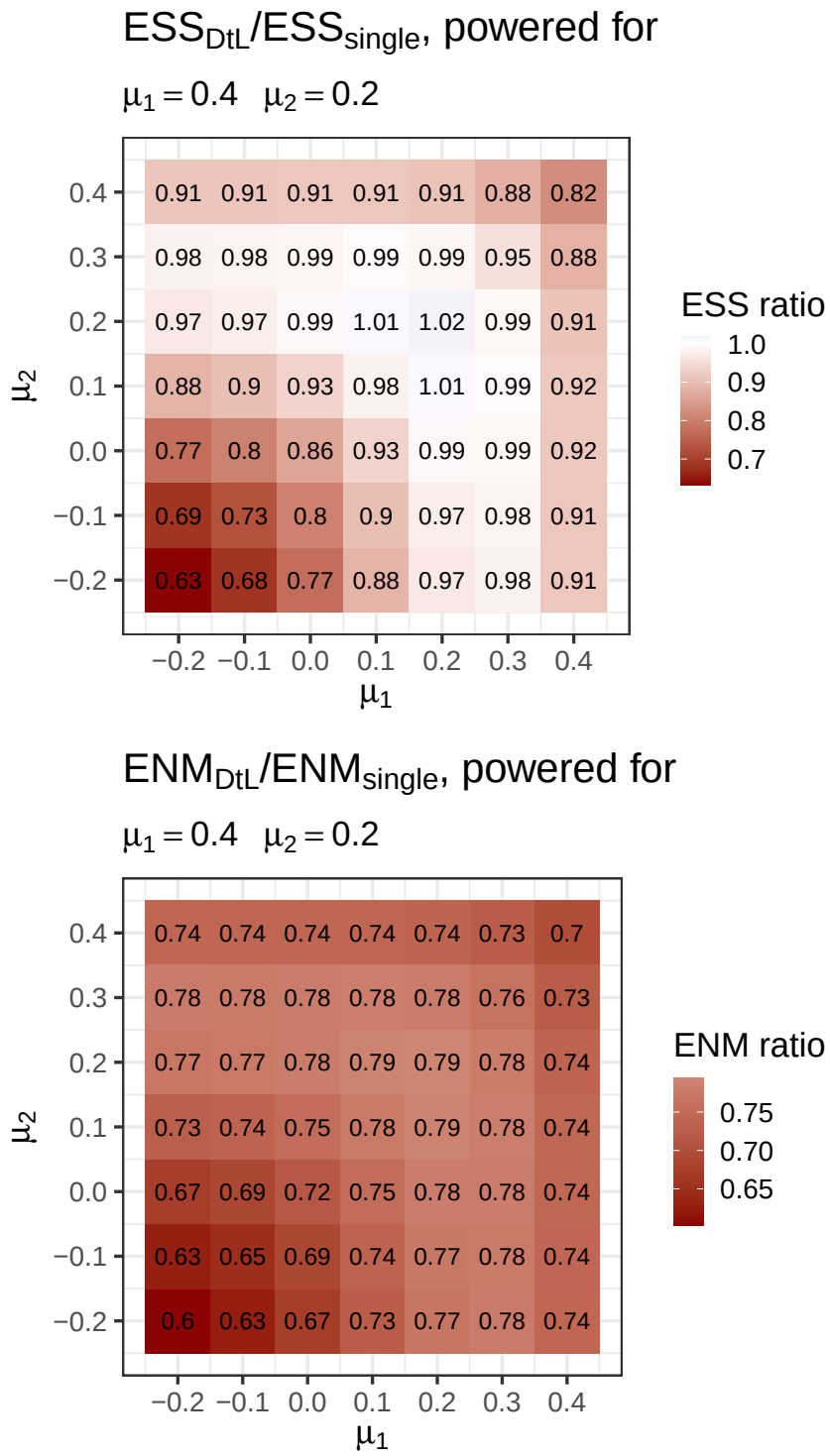


Fig. 4.10 Changes in ESS_{DtL}/ESS_{single} and ENM_{DtL}/ENM_{single} for fixed design with $\{K = 2, K_{max} = 1, m = 1\}$ and design is powered for outcome effects $\boldsymbol{\delta}_\beta = (0.4, 0.2)$.

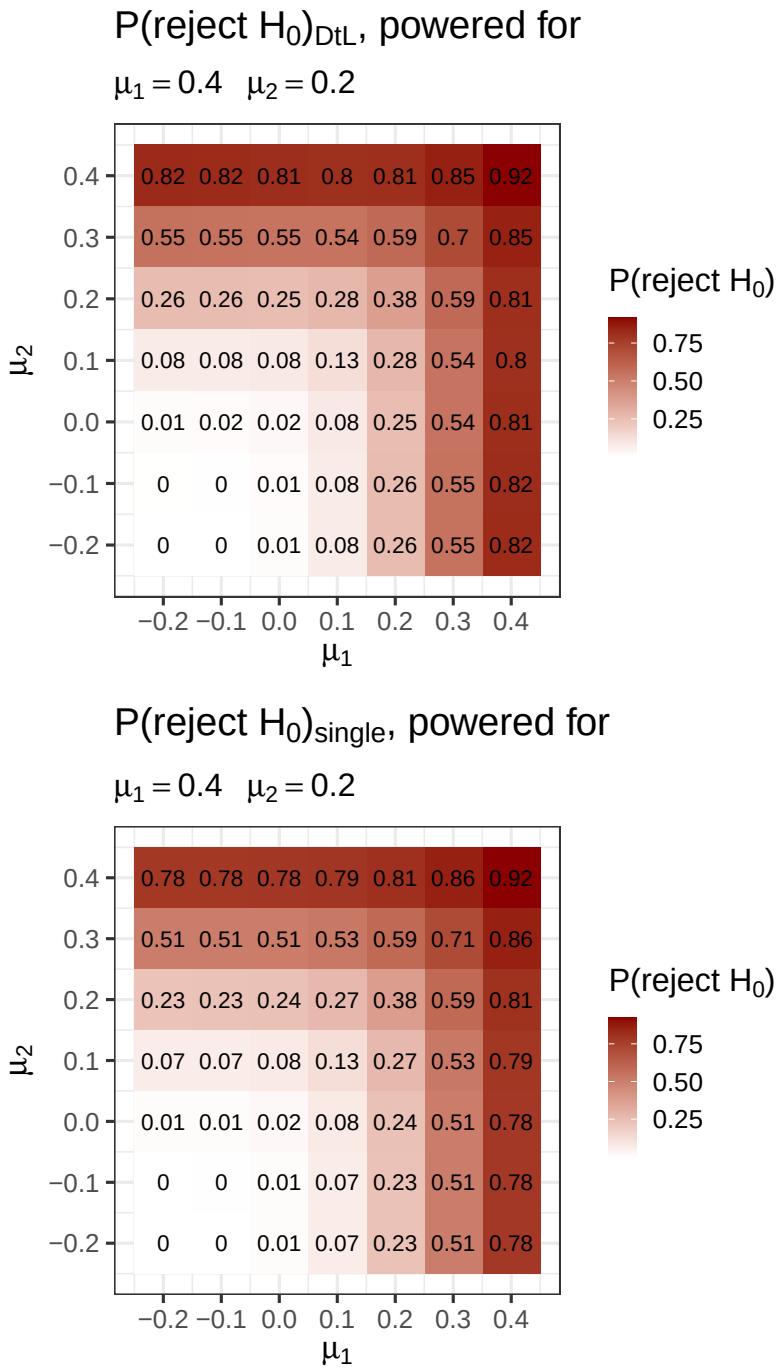


Fig. 4.11 Changes in the probability of rejecting H_0 for the DtL and single-stage designs with $\{K = 2, K_{max} = 1, m = 1\}$ and design is powered for outcome effects $\boldsymbol{\delta}_{\beta} = (0.4, 0.2)$.

4.6.2.2 Three outcomes

In the case of $K = 3, K_{max} = 1, m = 1$, probability of rejecting H_0, ESS ratio and ENM ratio are examined for a selection of true effect sizes $\{\mu_1, \mu_2, \mu_3\}$ in Table 4.3. The design realisations are $\{r = 2.435647, N = 72\}$ for the DtL design and $\{r = 2.380403, N = 59\}$ for the single-stage design. The results are in agreement with the $K = 2$ case: probability of rejecting H_0 is similar for both approaches for most featured cases and slightly lower for the single stage approach when one outcome has true effect as anticipated and the remaining outcomes have zero or harmful true effects. ESS ratio is greater than 1, i.e., favouring the single stage design, only when all outcomes have effects equal to δ_0. Again the ENM ratio is less than 1 in all cases.

μ_1	μ_2	μ_3	p(rej. H_0)$_{DtL}$	p(rej. H_0)$_{SS}$	$ESS_{DtL/SS}$	$ENM_{DtL/SS}$	Description
0.4	0.4	0.4	0.95	0.96	0.80	0.47	All outcomes have effect δ_1
0.4	0.2	0.2	0.81	0.80	0.95	0.52	Effects as anticipated (power)
0.4	0.0	0.0	0.82	0.76	0.97	0.53	One outcome has no effect
0.4	-0.2	-0.2	0.83	0.76	0.97	0.53	One outcome is harmful
0.0	0.0	0.0	0.02	0.02	0.96	0.52	Global null (type I error)
0.3	0.3	0.3	0.77	0.77	0.97	0.53	All have some effect $< \delta_1$
0.2	0.2	0.2	0.43	0.43	1.09	0.57	All outcomes have effect δ_0

Table 4.3 p(reject H_0| true effects $\boldsymbol{\mu}$), expected sample size ratios and expected number of measurements ratios for drop the loser design and single stage design, where $\{K = 3, K_{max} = 1, m = 1\}$ and design is powered for outcome effects $\boldsymbol{\delta}_\beta = (0.4, 0.2, 0.2)$. $ESS_{DtL/SS}$: ESS ratio. $ENM_{DtL/SS}$: ENM ratio.

4.7 Discussion

We have examined two approaches to generalising multi-outcome designs to allow trials that seek to determine if there exist some m of out K outcomes in a single treatment arm that show promise. Multiple primary outcome designs and co-primary outcome designs, in comparison, allow only $m = 1$ and $m = K$ respectively.

The first approach, a multi-outcome multi-stage design, was compared to a multi-stage design with a single composite outcome. As outcome correlation increases, the ESS and ENM of the proposed approach decrease relative to the composite approach, and were superior in all tested cases with correlation $\rho \geq 0.5$. When different true outcome effects are examined, ESS and ENM are generally lower for the proposed design, and are only greater than the composite design when more outcomes are efficacious than anticipated. The probability of rejecting the null hypothesis is more robust under the proposed approach, remaining close to nominal power when some true outcome effects are lower than anticipated while this probability decreases under the composite approach. Furthermore, rejecting the null hypothesis when no outcomes have the desired effect size is less likely using the proposed approach.

The second approach, a multi-outcome, two-stage DtL design, was compared to a single-stage design. Again the ESS and ENM of the proposed approach decrease compared to the existing approach as correlation increases. ENM was superior in the proposed approach for all cases examined, while ESS was superior in 42 out of 45 cases. Furthermore, in greater than 50% of cases, the ENM was reduced by at least half. When different true outcome effects were examined, ENM was reduced under the proposed DtL design compared to the single stage design in all cases, while ESS was reduced in 46 out of 49 cases. The probability of rejecting the null hypothesis when one outcome was as efficacious as anticipated while other outcomes had lower effect sizes than anticipated was similar for both approaches. However, the rejection probability was slightly greater than the required power for the DtL approach and slightly lower for the single stage approach.

The proposed approaches allow investigators to measure, at least initially, a range of outcomes while reducing the high costs that may be associated with such trials. Furthermore, these approaches offer novel flexibility in the area of multiple-outcome clinical trials, allowing investigators to specify any number of outcomes for which promise must be shown. This

is a novel generalisation of existing designs, which are special cases in comparison as they require promise to be shown on either all outcomes or one or more outcomes.

Chapter 5

Discussion

5.1 Summary

Phase II binary outcome trials are a critical aspect of drug development. However, with high failure rates and high costs, it is valuable to find ways of making correct decisions more quickly. Chapter 2 presented two single-arm designs created to improve binary outcome trials in this respect. Our main goal with regards to these designs was to improve upon existing designs in terms of the three optimality criteria ESS(p_0), ESS(p_1) and N. Secondary goals included finding design realisations without simulation, introducing stochastic curtailment for treatments that show promise and making any design search computationally viable.

Comparing our proposed designs to a number of existing designs, we found that the proposed designs were superior in almost all cases, whether considering either single optimality criteria or a weighted combination of multiple optimality criteria.

A number of the concepts used in the proposed single-arm designs would also be novel in the two-arm randomised setting, the gold standard in trial design. As such, we presented in Chapter 3 a two-arm design with some of the same goals as the proposed single-arm approaches. Again, our main aim was to improve upon existing designs in terms of multiple optimality criteria, in this case ESS(p_0, p_0), ESS(p_0, p_1) and N. As with the single-arm case,

we introduced stochastic curtailment for promising treatments, obtained design realisations without simulation and made the design search computationally viable. When compared to existing designs using a weighted combination of multiple optimality criteria, our proposed two-arm design was superior for p_0 values greater than or equal to 0.4, and second to only one existing design, Carsten and Chen [17], otherwise. However, the design of Carsten and Chen was found to be sensitive to deviations in the specified response rates, with a considerably greater probability of rejecting the null hypothesis than the proposed design in situations where the true treatment difference was smaller than desired.

Another aspect of clinical trials is the measurement of multiple outcomes, which is typical in clinical trials, for a number of reasons discussed in Chapter 4. Amongst clinical trial designs where multiple key outcomes are measured, designs are generally powered to identify when either at least one outcome shows promise or all measured outcomes show promise. Chapter 4 generalised this concept by presenting two multi-outcome designs, both of which were powered to find when at least some specified number of outcomes shows promise. One design permitted any number of stages, while the other was a two-stage design that permitted dropping outcomes that were performing poorly. Our main goal was to improve on existing designs by creating designs that meet the needs of investigators in ways that existing trials do not. In particular, both designs offer a generalised framework in terms of seeking a specified number of efficacious outcomes, and this framework is novel in a multi-stage setting.

Beyond this, the first multi-outcome design resulted in reduced ESS and ENM compared to a multi-outcome multi-stage composite outcome design when correlation is high ($\rho \geq 0.5$), while also being less sensitive to deviations to the anticipated effect sizes. The second design resulted in reduced ESS and ENM compated to a multi-outcome single-stage design in most cases, and again was less sensitive to deviations to the anticipated effect sizes.

5.2 Limitations

While all proposed methods performed well in the comparisons that have been made, there are limitations to their use. The two single-arm designs proposed in Chapter 2 find designs with similar operating characteristics. However, one design, the m-stage design, completes a design search more quickly than other design, the SC design, and this difference is approximately one order of magnitude. As a result, it is currently difficult to foresee a situation where the SC design can be recommended over the m-stage design. However, increases in computing power will render "slow" design searches viable over time, and choosing one approach over another in terms of computation time may be trivial in the near future.

The two-arm design proposed in Chapter 3 uses a randomised block design. This aspect of trial design is not novel. Nevertheless, in peer review one reviewer expressed concern that in single-centre trials, investigators may engage in selection bias by successfully "guessing" the arm to which the next participant will be assigned [112]. This concern is attributable to block randomisation itself rather than the proposed trial design approach, but may still be briefly addressed. In the first instance, we assume that any randomised study is double-blinded, that is, both participants and investigators do not know which treatment is which [44]. Selection bias may be further minimised by ensuring that the investigator responsible for selection does not take part in participant treatment assignment. Such steps may be taken independently of the design approach. Indeed, the CONSORT 2010 checklist of information to include when reporting a randomised trial includes "describing any steps taken to conceal the allocation sequence until interventions were assigned" [113]. A further step that may be taken is to vary block sizes within a trial, though this would require an extension of the current work and is beyond the scope of this paper. If a trial uses multiple centres and the randomised blocking is stratified by centre, then some imbalance may occur. Due to the typical size of phase II

trials in oncology, we recommend using randomised blocks in a centralised system, that is, not stratified by centre, which would ensure balance.

The proposed single- and two-arm designs both use sequential monitoring, which may be seen as a limitation. If a number of participant results emerge in quick succession then the interim analysis may not take place at the planned information fraction, increasing the sample size compared to the ESS. Such a possibility has not negatively affected the popularity of the Simon design, though we admit that sample size inflation is more likely as the number of decisions increases. The same issues exist with regards to delayed responses, that is, when recruitment rate is so great or endpoint length so long that not all participant results are available at the point where a decision is to be made regarding stopping or not stopping the trial [80]. However, as detailed in Chapter 2, recruitment is often slow in clinical trials, with a median recruitment rate of approximately one patient per centre per month. Moreover, in Chapter 2 we provide numerous examples of investigators making go and no go decisions as a result of continuous monitoring, even when the trial design was single- or two-stage. In our proposed designs, the frequency of monitoring can be specified at the trial design stage, to accommodate the practical needs of the investigators. A stopping boundary check should be undertaken as soon as results for each complete block are available. If this is somehow not possible and there exist excess results beyond a whole block, a stopping boundary check may still be undertaken using the results for participants whose results constitute completed blocks.

A separate limitation of the sequential monitoring is that, depending on the design realisation, it is possible that a trial may end with as few as two participants if block size two is chosen, which may be undesirable in some circumstances. However, among the set of five comparisons in Chapter 3, this did not occur for any of the four optimality criteria when block size two was used. Across these design realisations, the median minimum number of participants was found to be nine. Still, conservative investigators may prefer to either

use a larger block size or to begin with a single large block (e.g. a block of size sixteen) before switching to a smaller block size (e.g. blocks of size four), guaranteeing a minimum number of participants equal to the size of the first block. In the latter case, this would require augmenting the existing code in order to obtain the trial operating characteristics. The block sizes used in practice may differ from the planned trial design. In this case, the stopping boundaries could be reassessed taking this into account. Furthermore, stopping boundaries and conditional power could be re-estimated given the trial information so far. However, such extensions are beyond the scope of this thesis.

Both the proposed single- and two-arm designs may obtain design realisations that improve considerably upon existing designs in terms of ESS(p_0) and ESS(p_1). However, a limitation of the proposed designs is that the greatest improvements with respect to these criteria come in general at the expense of an increase in maximum sample size N. This is not unusual in adaptive design, and Wald's SPRT [14] provides an extreme example, providing low values of ESS(p_0) and ESS(p_1) coupled with an infinite N. Furthermore, any design that permits early stopping has uncertainty in the final sample size. This is of practical concern as sample size uncertainty results propagates uncertainty in contract length, recruitment targets, and ultimately, funding, though it is possible to ameliorate some negative effects of this uncertainty [114]. It is also possible to reduce sample size uncertainty itself at the design stage: when choosing a design realisation from a set of admissible designs, one may prioritise a low maximum sample size or even to minimise maximum sample size. This can be achieved by comparing design realisations using the loss function with a high weight on N, and software to do this has been created [75]. An investigator may choose the best design subject to the largest maximum sample size that they are willing to accept, where "best" means assigning weights to $ESS(p_0)$, $ESS(p_1)$ and maximum sample size.

The proposed multi-outcome designs have limitations regarding their generality. Both rely on continuous outcomes, rather than allowing other outcome types, such as binary or

ordinal, either on their own or in combination. The designs are single-arm, rather than two-arm. There is a single final rejection boundary shared between all outcomes, rather than permitting different boundaries for each outcome. The second design permits only two stages, in contrast to the multi-arm DtL design, where a general number of stages are accommodated.

Arguably, a limitation of this work in general is that it focuses solely on frequentist methods. Bayesian methods can be used for phase II clinical trial design [115–118] and are becoming more widely used over time [115, 119]. However, some Bayesian and frequentist designs are closely related conceptually, for example, the frequentist CP-based approach used in this thesis and the Bayesian predictive power approach [10]. Furthermore, in the context of binary outcome trials, Bayesian and frequentist designs can both be described in exactly the same way, that is, using vectors of lower and upper stopping boundaries $\mathbf{f}$ and $\mathbf{e}$. One main advantage of Bayesian methods in this context is the ability to incorporate prior information, for example, data from a previous phase I trial. However, not all Bayesian designs do so, and instead use an uninformative prior (or priors). In contrast, frequentist methods are deemed to discard such data, or at best use it in as a summary by, for example, using the data to inform a future choice of p_1. However, using the data in this way still has merit in the single- and two-arm designs we propose, for example, in a seamless phase I/II trial: our design searches result in a series of admissible designs, all of which satisfy given operating characteristics. Some designs may have low ESS when the response rate is low, while others may have low ESS when the response rate is high. With this in mind, investigators could specify merely the operating characteristics and design approach in a phase II protocol, allowing flexibility to choose a particular design realisation once phase I data has been obtained. Bayesian designs may be created with Bayesian operating characteristics in mind, which can be more intuitive (and thus easier to explain) to non-statisticians. However, Bayesian designs may be required to satisfy certain frequentist operating characteristics. In the case that a Bayesian

design both uses an uninformative prior and must satisfy some typical frequentist operating characteristics, the resulting design realisation may confer no advantage over an equivalent frequentist design. Another advantage of some Bayesian designs is that they are more flexible in terms of allowing interim analyses to occur at points that are not fixed in advance, with only minor negative consequences in terms of Bayesian operating characteristics [118]. In one example of such a design, Lee and Liu [118] investigate the effect of this flexibility by assuming that the point at which the interim analyses will take place is random. However, in a clinical trial, allowing such flexibility could result in unconscious bias, with investigators able to undertake an interim analysis after observing a succession of positive results or avoiding an interim analysis after observing a succession of negative results. Finally, a fundamental difference between the frequentist and Bayesian frameworks is that the Bayesian framework must rely on simulation, with results subject to simulation error. In contrast, some frequentist approaches, such as those presented in Chapters 2 and 3, do not require simulation and are free from simulation error.

5.3 Recommendations

While the m-stage design performed well in all circumstances examined, other designs performed similarly when sole importance was placed on minimising maximum sample size N, disregarding ESS(p_0) and ESS(p_1). In particular, in one of three scenarios, the design of Mander and Thompson [8] achieved a better maximum sample size than the proposed designs. As such, existing designs should be preferred over the proposed designs when performance in similar for the optimality criterion of prime importance, and the existing design uses fewer interim analyses.

The proposed two-arm design performed better than existing designs when $p_0 \geq 0.4$, and we recommend its use in these circumstances. The design also outperforms existing designs when $p_0 = 0.3$ and the weighting of optimality criteria prioritises ESS(p_0) or ESS(p_1), either

on their own or in combination. We recommend our proposed design in these situations. Otherwise, the design of Carsten and Chen [17] achieves better operating characteristics. However, our proposed design is less sensitive to deviations from the anticipated response rates, and so we also recommend our design when there is at least moderate uncertainty regarding anticipated response rates. The Carsten and Chen design also requires results to be analysed after every pair of participants, compared to out flexible degree of monitoring, and so we also recommend our design when less frequent monitoring is desired.

In practical terms, the single- and two-arm designs can be used by calling the corresponding functions in R [74], after installing the `curtailment` package in R [75]. One function undertakes a single-arm design search, while another function undertakes a two-arm design search, given the appropriate requirements, for example, type-I error-rate, power and response rates. Admissible design realisations are returned, if they exist. A second function (for each design approach) takes as its input any chosen design realisation and returns the corresponding stopping boundaries. While the designs as proposed may be seen as complex, the final output is simply a collection of stopping boundaries. Following the design could be made as simple as checking a diagram similar to those in Figure 2.1. Providing ways to make a novel design more easy to understand may be crucial to the design being adopted for more widespread use [120].

The m-stage design is being considered for use in the upcoming single-arm Phase II trial Positioning Imatinib for Pulmonary Arterial Hypertension (PIPAH) [121]. Pulmonary arterial hypertension is a rare condition, with observed prevalence of 5-52 cases per million [122], and so using a trial design that can come to conclusion quickly would be beneficial.

The proposed multi-outcome designs would be of value for any investigator who seeks to conduct a multi-outcome trial that is powered to identify when a specified number of the (continuous) outcomes show promise. In particular, the multi-outcome multi-stage design shows improvements in ESS and ENM compared to a multi-stage composite design

when correlation between outcomes is high ($\rho \geq 0.5$). The multi-outcome DtL design also improves ESS and ENM when correlation is high, compared to the multi-outcome single-stage design, and so both designs are recommended when outcome correlation is (or is anticipated to be) high. The proposed designs can also be recommended when creating a composite outcome is not appropriate. Both proposed designs seek to reduce ENM, the multi-outcome multi-stage design by providing multiple interim analyses at which points the trial may stop for either a go or no go decision, the multi-outcome DtL design by allowing measurement of poorly performing outcomes to cease as well as allowing the trial to stop at a single interim analysis. As such, we recommend these designs when the cost of outcome measurement is high. Both designs showed less sensitivity to their comparators when the true effect sizes deviated from the anticipated effect sizes. As such, like the proposed two-arm outcome binary design above, we recommend the proposed multi-outcome designs when there is uncertainty regarding the anticipated effect sizes.

Both multi-outcome design approaches find design realisations using simulation, and as such simply report a single design realisation, again by calling a single function in the R package `moms` [76]. The multi-outcome multi-stage design finds a single design realisation and reports the stopping boundaries for each stage. These are found using the equation by Wang-Tsiatis [105] (Section 4.3). With the stopping boundaries known, the investigator will end the trial only if m upper boundaries or $K-m+1$ lower boundaries are crossed simultaneously.

The stopping boundaries for the multi-outcome DtL design can be found by inverting the two-stage case of Jennison and Turnbull's equation for CP [10], giving

$$f_k = \frac{1}{\sqrt{\mathscr{I}_{1k}}}\left[\sqrt{\mathscr{I}_{2k}-\mathscr{I}_{1k}}\,\Phi^{-1}(CP_L)+Z_\alpha\sqrt{\mathscr{I}_{2k}}-(\mathscr{I}_{2k}-\mathscr{I}_{1k})\delta_{1k}\right]$$
$$e_k = \frac{1}{\sqrt{\mathscr{I}_{1k}}}\left[\sqrt{\mathscr{I}_{2k}-\mathscr{I}_{1k}}\,\Phi^{-1}(CP_U)+Z_\alpha\sqrt{\mathscr{I}_{2k}}-(\mathscr{I}_{2k}-\mathscr{I}_{1k})\delta_{1k}\right],$$

where $\mathscr{I}_{1k}, \mathscr{I}_{2k}$ are the outcome-specific information $\mathscr{I}$ for each stage. As above, the investigator stops the trial at the interim if m upper boundaries or $K-m+1$ lower boundaries are crossed simultaneously. Otherwise, some outcomes are dropped at the interim and the trial continues. However, the number of outcomes retained is $\min\left(K_{max}, \sum_{k=1}^{K} \mathbb{I}\left(CP_k(\delta_{1k}) > CP_L\right)\right)$. Consequently, if the number of outcomes to be dropped is greater than the number of outcomes that are lower than the lower boundary, the investigator must be able to know which outcomes to be dropped. That is, they must know the CP of the outcomes at the interim. This is taken care of using a function in the R package `moms`, where the interim test statistics can be entered and the appropriate decision (to stop and reject H_0, stop and not reject H_0 or continue and retain certain outcomes) is given, in addition to the CP of each outcome [76].

5.4 Future work

The proposed single-arm designs could be generalised in a number of ways in future work. We have assumed that all participants' results are available before any subsequent enrollment. It would be valuable to investigate the effects of delayed responses on ESS in curtailed designs, for a range of recruitment rates and endpoint lengths. It may be particularly worthwhile to consider how to proceed when a decision to stop has been made just before observing further results. There may be, for example, a pause in recruitment, then the initial stopping decision may be finalised or overturned. Such further work could quantify to what extent delayed responses increase ESS in designs that use curtailment.

In case of a desire to collect a certain degree of information in a trial, a trial could be specified to end only after data is available for some minimum number of participants. With regard to estimation, estimates of confidence intervals and p values could be compared to those from existing design types.

The effect of less frequent monitoring on curtailed designs could be examined further in terms of, for example, how to optimise monitoring frequency between improved operating characteristics, the perceived cost of each interim analysis and bias.

Examples were given of clinical trials that informally used continuous monitoring to stop early (Section 1.1.3.1). However, it is difficult to identify informal continuous monitoring in trials where no early stopping took place. Consequently, it may be illuminating to attempt to quantify what proportion of trials have used continuous monitoring and how frequently monitoring takes place within such trials. This may be undertaken through a survey or similar.

Some of the suggested future work for single-arm designs also apply to two-arm designs: the effect of delayed responses on ESS; permitting a trial to end only after some minimum number of participant results, and the effect of less frequent monitoring and how to optimise this. For the proposed two-arm design, stopping decisions are made after every block of participants, where a block must contain at least two participants, while future work could consider how to find stopping boundaries that would be appropriate for after every single participant, for a design that uses SC.

As discussed in Chapter 2, when using any clinical trial design that permits early stopping, the maximum likelihood estimator may be biased. Estimators have been developed that can be used for inference in trials with more than two stages [82, 123]. In particular, Bibbona and Rubba [123] present an estimator for multi-stage multinomial clinical trials. This estimator may be able to be applied to two-arm multi-stage trials.

For both the proposed single- and two-arm designs, we could further investigate quantiles of sample size compared to other designs. In this thesis, single-arm designs were compared to Simon's design, but a comprehensive comparison could involve both more designs and comparisons between two-arm designs. For example, Hanfelt et al. [124] modified Simon's design to optimise for median sample size.

There is scope for future work regarding the proposed multi-outcome designs. This may involve generalising the number of stages in the DtL design, which could result in further savings in ESS and ENM. The proposed approaches are for a single-arm trial, while Sozu et al. use a two-arm design [49]. Extending these designs to two arms would give investigators more options with regards to trial design. Other possible generalisations include allowing interim boundaries and CP boundaries to differ across outcomes for the multi-outcome and DtL approaches respectively and allowing final boundaries to differ, for both approaches. Such lack of generalisation may be considered limitations of the proposed approaches. For the DtL design, it would be worthwhile to undertake a sensitivity analysis to fully explore the effects of varying the interim CP bounds.

It may be possible to divide outcomes into those that must show promise, effectively a subset containing multiple co-primary outcomes, and those among which only a subset are required to show promise. This would be of use if, for example, a treatment is required to show an effect on some safety outcome and simultaneously show an effect on some specified number of efficacy outcomes. Other possible extensions include the introduction of alpha spending, rather than using an overall type-I error-rate, and extending to other types of outcome, for example binary outcomes.

5.5 Conclusion

The clinical trial designs proposed in this thesis make novel contributions to the literature, both in binary outcome trial design and in continuous multiple outcome trial design. The designs offer considerably improved operating characteristics compared to existing designs. Particularly with regard to the proposed binary outcome designs, the designs are simple for investigators to use. Widespread use of these designs would speed up drug development.

References

[1] C.H. Wong, K.W. Siah, and A.W. Lo. Estimation of clinical trial success rates and related parameters. Biostatistics, 20(2):273–286, 2019.

[2] L. Martin, M. Hutchens, C. Hawkins, and A. Radnov. How much do clinical trials cost? Nature Reviews Drug Discovery, 16:381–382, 2017.

[3] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). European Journal of Cancer, 45(2):228–247, 2009. Response assessment in solid tumours (RECIST): Version 1.1 and supporting papers.

[4] R.P. A'Hern. Sample size tables for exact single-stage phase II designs. Statistics in Medicine, 20(6):859–866, 2001.

[5] R. Simon. Optimal two-stage designs for phase II clinical trials. Controlled Clinical Trials, 10:1–10, 1989.

[6] T. Jaki. Uptake of novel statistical methods for early-phase clinical studies in the UK public sector. Clinical Trials, 10:344–346, 2013.

[7] A. Ivanova, B. Paul, O. Marchenko, G. Song, N. Patel, and S.J. Moschos. Nine-year change in statistical design, profile, and success rates of Phase II oncology trials. Journal of Biopharmaceutical Statistics, 26(1):141–149, 2016.

[8] A.P. Mander and S.G. Thompson. Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. Contemporary Clinical Trials, 31(6):572 – 578, 2010.

[9] Y. Chi and C. Chen. Curtailed two-stage designs in phase II clinical trials. Statistics in Medicine, 27:6175–6189, 2008.

[10] C. Jennison and B.W. Turnbull. Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC, 2000.

[11] K. Kunzmann, M.J. Grayling, K.M. Lee, D.S. Robertson, K. Rufibach, and J.M.S. Wason. Conditional power and friends: The why and how of (un)planned, unblinded sample size recalculations in confirmatory trials. arXiv preprint arXiv:2010.06567, 2020.

[12] A.O. Ayanlowo and D.T. Redden. Stochastically curtailed phase II clinical trials. Statistics in Medicine, 26(7):1462–1472, 2007.

[13] C.U. Kunz and M. Kieser. Curtailment in single-arm two-stage phase II oncology trials. Biometrical Journal, 54(4):445–456, 2012.

[14] A. Wald. Sequential Analysis. Dover, 1947.

[15] J. Whitehead. The Design and Analysis of Sequential Clinical Trials. Wiley, 1997.

[16] C.M. Chen, Y. Chi, and H.M. Chang. Curtailed two-stage design for comparing two arms in randomized phase II clinical trials. Journal of Biopharmaceutical Statistics, 28(5):939–950, 2018.

[17] C. Carsten and P. Chen. Curtailed two-stage matched pairs design in double-arm Phase II clinical trials. Journal of Biopharmaceutical Statistics, 26(5):816–822, 2016.

[18] J.M.S. Wason, P. Brocklehurst, and C. Yap. When to keep it simple–adaptive designs are not always useful. BMC Medicine, 17(1):1–7, 2019.

[19] M.K. Campbell, C. Snowdon, D. Francis, D. Elbourne, A.M. Mcdonald, R. Knight, V. Entwistle, J. Garcia, I. Roberts, and A. Grant. Recruitment to randomised trials: strategies for trial enrolment and participation study. Health Technology Assessment, 11(48), 2007.

[20] S.J. Walters, I. Bonacho, O. Bortolami, L. Flight, D. Hind, R.M. Jacques, C. Knox, B. Nadin, J. Rothwell, M. Surtees, and S.A. Julious. Recruitment and retention of participants in randomised controlled trials : a review of trials funded and published by the United Kingdom Health Technology Assessment Programme. BMJ Open, pages 1–10, 2017.

[21] J.A. Todd, M. Evangelou, A.J. Cutler, M.L. Pekalski, M. Walker, H.E. Stevens, L. Porter, D.J. Smyth, D.B. Rainbow, R.C. Ferreira, L. Esposito, K.M.D. Hunter, K. Loudon, K. Irons, J.H. Yang, C.J.M. Bell, H. Schuilenburg, J. Heywood, B. Challis, S. Neupane, P. Clarke, G. Coleman, S. Dawson, D. Goymer, K. Anselmiova, J. Kennet, J. Brown, S.L. Caddy, J. Lu, J. Greatorex, I. Goodfellow, C. Wallace, T.I. Tree, M. Evans, A.P. Mander, S. Bond, L.S. Wicker, and F. Waldron-lynch. Regulatory T cell responses in participants with type 1 diabetes after a single dose of interleukin-2 : A non-randomised , open label , adaptive dose-finding trial. PLOS Medicine, pages 1–33, 2016.

[22] L. McCabe, I.R. White, N.V. Vinh Chau, E. Barnes, S.L. Pett, G.S. Cooke, and A.S. Walker. The design and statistical aspects of VIETNARMS: a strategic post-licensing trial of multiple oral direct-acting antiviral hepatitis C treatment strategies in Vietnam. Trials, 21:1–12, 2020.

[23] T.A. Santana, F.M. Cruz, D.C. Trufelli, J. Glasberg, and A. Del Giglio. Carbamazepine for prevention of chemotherapy-induced nausea and vomiting: a pilot study. Sao Paulo Medical Journal, 132:147 – 151, 00 2014.

[24] A. Necchi, P. Giannatempo, L. Mariani, E. Farè, D. Raggi, M. Pennati, N. Zaffaroni, F. Crippa, A. Marchianò, N. Nicolai, M. Maffezzini, E. Togliardi, M.G. Daidone, A.M. Gianni, R. Salvioni, and F. De Braud. PF-03446962, a fully-human monoclonal antibody against transforming growth-factor β (TGFβ) receptor ALK1, in pre-treated patients with urothelial cancer: an open label, single-group, phase 2 trial. Investigational New Drugs, 32(3):555–560, Jun 2014.

[25] M. Mego, D. Svetlovska, V. Miskovska, J. Obertova, P. Palacka, J. Rajec, Z. Sycova-Mila, M. Chovanec, K. Rejlekova, P. Zuzak, D. Ondrus, S. Spanik, M. Reckova, and J. Mardiak. Phase II study of everolimus in refractory testicular germ cell tumors. Urologic Oncology, 34(3):17–22, Mar 2016.

[26] L.M. Wagner, M. Fouladi, A. Ahmed, M.D. Krailo, B. Weigel, S.G. DuBois, L.A. Doyle, H. Chen, and S.M. Blaney. Phase II study of cixutumumab in combination with temsirolimus in pediatric patients and young adults with recurrent or refractory sarcoma: a report from the Children's Oncology Group. Pediatric Blood & Cancer, 62(3):440–444, Mar 2015.

[27] S.M. Stein, A. Tiersten, H.S. Hochster, S.V. Blank, B. Pothuri, J. Curtin, I. Shapira, B. Levinson, P. Ivy, B. Joseph, A.K. Guddati, and F. Muggia. A phase 2 study of oxaliplatin combined with continuous infusion topotecan for patients with previously treated ovarian cancer. International Journal of Gynecologic Cancer, 23(9):1577–1582, 2013.

[28] S.S. Yu, K. Athreya, S.V. Liu, A.V. Schally, D. Tsao-Wei, S. Groshen, D.I. Quinn, T.B. Dorff, S. Xiong, J. Engel, and J. Pinski. A phase II trial of AEZS-108 in castration- and taxane-resistant prostate cancer. Clinical Genitourinary Cancer, 15(6):742 – 749, 2017.

[29] A.J. Moskowitz, P.A. Hamlin, M. Perales, J. Gerecitano, S.M. Horwitz, M.J. Matasar, A. Noy, M.L. Palomba, C.S. Portlock, D.J. Straus, T. Graustein, A.D. Zelenetz, and C.H. Moskowitz. Phase II study of bendamustine in relapsed and refractory hodgkin lymphoma. Journal of Clinical Oncology, 31(4):456–460, 2013. PMID: 23248254.

[30] H.H. Yoon, N.R. Foster, J.P. Meyers, P.D. Steen, D.W. Visscher, R. Pillai, D.M. Prow, C.M. Reynolds, B.T. Marchello, R.B. Mowat, B.I. Mattar, C. Erlichman, and M.P. Goetz. Gene expression profiling identifies responsive patients with cancer of unknown primary treated with carboplatin, paclitaxel, and everolimus: NCCTG N0871 (alliance). Annals of Oncology, 27(2):339 – 344, 2016.

[31] J.M. Sepulveda-Sanchez, M.A. Vaz, C. Balana, M. Gil-Gil, G. Reynes, O. Gallego, M. Martinez-Garcia, E. Vicente, M. Quindos, R. Luque, A. Ramos, Y. Ruano, P. Perez-Segura, M. Benavides, P. Sanchez-Gomez, and A. Hernandez-Lain. Phase II trial of dacomitinib, a pan-human EGFR tyrosine kinase inhibitor, in recurrent glioblastoma patients with EGFR amplification. Neuro-Oncology, 19(11):1522–1531, Oct 2017.

[32] K.S. Pedersen, G.P. Kim, N.R. Foster, A. Wang-Gillam, C. Erlichman, and R.R. McWilliams. Phase II trial of gemcitabine and tanespimycin (17AAG) in metastatic pancreatic cancer: a Mayo Clinic phase II consortium study. Investigational New Drugs, 33(4):963–968, Aug 2015.

[33] Y. Odia, T.N. Kreisl, D. Aregawi, E.K. Innis, and H.A. Fine. A phase II trial of tamoxifen and bortezomib in patients with recurrent malignant gliomas. Journal of Neuro-Oncology, 125(1):191–195, Oct 2015.

[34] J. Whitehead. On the bias of maximum likelihood estimation following a sequential test. Biometrika, 73(3):573–581, 1986.

[35] E.N. Atkinson and B.W. Brown. Confidence limits for probability of response in multistage phase II clinical trials. Biometrics, 41(3):741–744, 1985.

[36] S.H. Jung, T. Lee, K.M. Kim, and S.L. George. Admissible two-stage designs for phase II cancer clinical trials. Statistics in Medicine, 23:561–569, 2004.

[37] A.P. Mander, J.M.S. Wason, M.J. Sweeting, and S.G. Thompson. Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. Pharmaceutical Statistics, 11(2):91–96, 2012.

[38] A.J. Vickers, V. Ballen, and H.I. Scher. Setting the bar in phase II trials: the use of historical data for determining "go/no go" decision for definitive phase III testing. Clinical Cancer Research, 13(3):972–6, feb 2007.

[39] M.J. Grayling, M. Dimairo, A.P. Mander, and T.F. Jaki. A review of perspectives on the use of randomization in phase II oncology trials. JNCI: Journal of the National Cancer Institute, jun 2019.

[40] L. Rubinstein, J. Crowley, P. Ivy, M. LeBlanc, and D. Sargent. Randomized phase II designs. Clinical Cancer Research, 15(6):1883–1890, 2009.

[41] Prasad V. and A. Oseran. Do we need randomised trials for rare cancers? European Journal of Cancer, 51(11):1355 – 1357, 2015.

[42] F. Lasch, K. Weber, M.M. Chao, and A. Koch. A plea to provide best evidence in trials under sample-size restrictions: the example of pioglitazone to resolve leukoplakia and erythroplakia in fanconi anemia patients. Orphanet Journal of Rare Diseases, 12(102), 2017.

[43] M.J. Grayling and A.P. Mander. Do single-arm trials have a role in drug development plans incorporating randomised trials? Pharmaceutical Statistics, 15(2):143–151, mar 2016.

[44] D.G. Altman. Practical Statistics for Medical Research. Chapman and Hall/CRC, 1990.

[45] J. Langrand-Escure, R. Rivoirard, M. Oriol, F. Tinquaut, C. Rancoule, F. Chauvin, N. Magné, and A. Bourmaud. Quality of reporting in oncology phase II trials: A 5-year assessment through systematic review. PloS One, 12(12):e0185536, 2017.

[46] S. Jung. Randomized phase II trials with a prospective control. Statistics in Medicine, 27(4):568–583, feb 2008.

[47] B.R. Celli, W. MacNee, A. Agusti, A. Anzueto, B. Berg, A.S. Buist, P.M.A. Calverley, N. Chavannes, T. Dillard, B. Fahy, A. Fein, J. Heffner, S. Lareau, P. Meek, F. Martinez, W. McNicholas, J. Muris, E. Austegard, R. Pauwels, S. Rennard, A. Rossi, N. Siafakas, B. Tiep, J. Vestbo, E. Wouters, and R. ZuWallack. Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. European Respiratory Journal, 23(6):932–946, 2004.

[48] P.R. Williamson, D.G. Altman, H. Bagley, K.L. Barnes, J.M. Blazeby, S.T. Brookes, M. Clarke, E. Gargon, S. Gorst, N. Harman, J.J. Kirkham, A. McNair, C.A.C. Prinsen, Jochen S., C.B. Terwee, and B. Young. The COMET handbook: version 1.0. Trials, 18(S3), jun 2017.

[49] T. Sozu, T. Sugimoto, T. Hamasaki, and S.R. Evans. Sample Size Determination in Clinical Trials with Multiple Endpoints. Springer, 2015.

[50] W. Maurer and F. Bretz. Memory and other properties of multiple test procedures generated by entangledgraphs. Statistics in Medicine, 32(10):1739–1753, 2013.

[51] J. Wason, N. Stallard, J. Bowden, and C. Jennison. A multi-stage drop-the-losers design for multi-arm clinical trials. Statistical Methods in Medical Research, 26(1):508–524, 2017.

[52] T. Hamasaki, S.R. Evans, and K. Asakura. Design, data monitoring, and analysis of clinical trials with co-primary endpoints: A review. Journal of Biopharmaceutical Statistics, 28(1):28–51, oct 2017.

[53] Y. Ando and T. Hamasaki. Practical issues and lessons learned from multi-regional clinical trials via case examples: a Japanese perspective. Pharmaceutical Statistics, 9(3):190–200, 2010.

[54] Y. Ando, T. Hamasaki, S.R. Evans, K. Asakura, T. Sugimoto, T. Sozu, and Y. Ohno. Sample size considerations in clinical trials when comparing two interventions using multiple co-primary binary relative risk contrasts. Statistics in Biopharmaceutical Research, 7(2):81–94, 2015.

[55] K. Asakura, T. Hamasaki, T. Sugimoto, K. Hayashi, S.R Evans, and T. Sozu. Sample size determination in group-sequential clinical trials with two co-primary endpoints. Statistics in Medicine, 33(17):2897–2913, 2014.

[56] Y. Cheng, S. Ray, M. Chang, and S. Menon. Statistical monitoring of clinical trials with multiple co-primary endpoints using multivariate B-value. Statistics in Biopharmaceutical Research, 6(3):241–250, 2014.

[57] T. Hamasaki, K. Asakura, S.R. Evans, T. Sugimoto, and T. Sozu. Group-sequential strategies in clinical trials with multiple co-primary outcomes. Statistics in Biopharmaceutical Research, 7(1):36–54, 2015.

[58] R.J. Cook and V.T. Farewell. Guidelines for monitoring efficacy and toxicity responses in clinical trials. Biometrics, pages 1146–1152, 1994.

[59] C. Jennison and B.W. Turnbull. Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. Biometrics, pages 741–752, 1993.

[60] S. Schüler, M. Kieser, and G. Rauch. Choice of futility boundaries for group sequential designs with two endpoints. BMC Medical Research Methodology, 17(1):1–10, 2017.

[61] K. Asakura, T. Hamasaki, and S.R. Evans. Interim evaluation of efficacy or futility in group-sequential trials with multiple co-primary endpoints. Biometrical Journal, 59(4):703–731, 2017.

[62] M.R. Conaway and G.R. Petroni. Bivariate Sequential Designs for Phase II Trials. Biometrics, 51(2):656, jun 1995.

[63] M.R. Conaway and G.R. Petroni. Designs for Phase II Trials Allowing for a Trade-Off between Response and Toxicity. Biometrics, 52(4):1375, 1996.

[64] R. Ristl, D. Xi, E. Glimm, and M. Posch. Optimal exact tests for multiple binary endpoints. Computational Statistics and Data Analysis, 122:1–17, 2018.

[65] T. Jaki and L.V. Hampson. Designing multi-arm multi-stage clinical trials using a risk-benefit criterion for treatment selection. Statistics in Medicine, 35(4):522–533, 2016.

[66] P.F. Thall and S.C. Cheng. Optimal two-stage designs for clinical trials based on safety and efficacy. Statistics in Medicine, 20(7):1023–1032, 2001.

[67] J. Bryant and R. Day. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics, pages 1372–1383, 1995.

[68] T. Hamasaki, K. Asakura, S.R. Evans, and T. Ochiai. Group-sequential clinical trials with multiple co-objectives. Springer, 2016.

[69] S. Urach and M. Posch. Multi-arm group sequential designs with a simultaneous stopping rule. Statistics in Medicine, 35(30):5536–5550, 2016.

[70] J.M.S. Wason, D. Magirr, M. Law, and T. Jaki. Some recommendations for multi-arm multi-stage trials. Statistical Methods in Medical Research, 25(2):716–727, 2016.

[71] M.J. Grayling, J.M.S. Wason, and A.P. Mander. Efficient determination of optimised multi-arm multi-stage experimental designs with control of generalised error-rates. arXiv preprint arXiv:1712.00229, 2017.

[72] P. Delorme, P.L. de Micheaux, B. Liquet, and J. Riou. Type-II generalized family-wise error rate formulas with application to sample size determination. Statistics in Medicine, 35(16):2687–2714, 2016.

[73] J. Mielke, B. Jones, M. Posch, and F. König. Testing Procedures for Claiming Success on at Least k Out of m Hypotheses with an Application to Biosimilar Development. Statistics in Biopharmaceutical Research, 13(1):106–112, 2021.

[74] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[75] M. Law. Github repository. https://github.com/martinlaw/curtailment, 2021.

[76] M. Law. Github repository. https://github.com/martinlaw/moms, 2021.

[77] M. Law, M.J. Grayling, and A.P. Mander. Optimal curtailed designs for single arm phase II clinical trials. arXiv preprint arXiv:1909.03017, Sep 2019.

[78] G. Shan, J.J. Chen, and C. Ma. Boundary problem in Simon's two-stage clinical trial designs. Journal of Biopharmaceutical Statistics, 27(1):25–33, 2017.

[79] M. R. Sharma, K. Wroblewski, B. N. Polite, J. A. Knost, J. A. Wallace, S. Modi, B. G. Sleckman, D. Taber, E. E. Vokes, W. M. Stadler, and H. L. Kindler. Dasatinib in previously treated metastatic colorectal cancer: a phase II trial of the University of Chicago Phase II Consortium. Investigational New Drugs, 30(3):1211–5, 2012.

[80] L.V. Hampson and C. Jennison. Group sequential tests for delayed responses (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(1):3–54, 2013.

[81] A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. Annals of Mathematical Statistics, 19(3):326–339, 09 1948.

[82] M.A. Girshick, F. Mosteller, and L.J. Savage. Unbiased estimates for certain binomial sampling problems with applications. Annals of Mathematical Statistics, 17(1):13–23, 03 1946.

[83] S.H. Jung and K.M. Kim. On the estimation of the binomial probability in multistage clinical trials. Statistics in Medicine, 23(6):881–896, 2004.

[84] M. Chang, H. Wieand, and V. Chang. The bias of the sample proportion following a group sequential phase II clinical trial. Statistics in Medicine, 8(5):563–570, 1989.

[85] H.Y. Guo and A. Liu. A simple and efficient bias-reduced estimator of response probability following a group sequential phase II trial. Journal of Biopharmaceutical Statistics, 15(5):773–781, 2005.

[86] T. Koyama and H. Chen. Proper inference from Simon's two-stage designs. Statistics in Medicine, 27(16):3145–54, 2008.

[87] M.J. Grayling. singlearm. https://github.com/mjg211, 2019.

[88] C.U. Kunz and M. Kieser. Simon's minimax and optimal and Jung's admissible two-stage designs with or without curtailment. Stata Journal, 11(2):240–254(15), 2011.

[89] M. Law, M.J. Grayling, and A.P. Mander. A stochastically curtailed two-arm randomised phase II trial design for binary outcomes. Pharmaceutical Statistics, 2020.

[90] S. Roychoudhury, N. Scheuer, and B. Neuenschwander. Beyond p-values: A phase II dual-criterion design with statistical significance and clinical relevance. Clinical Trials, 15(5):452–461, 2018.

[91] P. Frewer, P. Mitchell, C. Watkins, and J. Matcham. Decision-making in early clinical drug development. Pharmaceutical Statistics, 15(3):255–263, 2016.

[92] R. Fisch, I. Jones, J. Jones, J. Kerman, G.K. Rosenkranz, and H. Schmidli. Bayesian design of proof-of-concept trials. Therapeutic Innovation & Regulatory Science, 49(1):155–162, 2014.

[93] G. Shan. Comments on 'two-sample binary phase 2 trials with low type i error and low sample size'. Statistics in Medicine, 36(21):3437–3438, 2017.

[94] J.L. Kepner. On group sequential designs comparing two binomial proportions. Journal of Biopharmaceutical Statistics, 20(1):145–159, 2010.

[95] W. Hou, M.N. Chang, S. Jung, and Y. Li. Designs for randomized phase II clinical trials with two treatment arms. Statistics in Medicine, 32(25):4367–4379, 2013.

[96] G. Shan, C. Ma, A.D. Hutson, and G.E. Wilding. Randomized two-stage phase II clinical trial designs based on Barnard's exact test. Journal of Biopharmaceutical Statistics, 23(5):1081–1090, 2013.

[97] S. Jung and D.J. Sargent. Randomized Phase II clinical trials. Journal of Biopharmaceutical Statistics, 24(4):802–816, 2014.

[98] M. Cellamare and V. Sambucini. A randomized two-stage design for phase II clinical trials based on a Bayesian predictive approach. Statistics in Medicine, 34(6):1059–1078, 2015.

[99] S. Litwin, S. Basickes, and E.A. Ross. Two-sample binary phase 2 trials with low type I error and low sample size. Statistics in Medicine, 36(9):1383–1394, 2017.

[100] M.J. Grayling. ph2rand. https://github.com/mjg211, 2019.

[101] K.A. Blum, J.L. Johnson, S. Jung, B.D. Cheson, and N.L. Bartlett. Serious pulmonary toxicity with SGN-30 and gemcitabine, vinorelbine, and liposomal doxorubicin in patients with relapsed/refractory hodgkin lymphoma (HL): Cancer and leukemia group B (CALGB) 50502. Blood, 112(11):232–232, 2008.

[102] Wassmer G. and F. Pahlke. rpact: Confirmatory adaptive clinical trial design and analysis. https://CRAN.R-project.org/package=rpact, 2019. R package version 2.0.6.

[103] S. Schüler, M. Kieser, and G. Rauch. Choice of futility boundaries for group sequential designs with two endpoints. BMC Medical Research Methodology, 17(1):119, dec 2017.

[104] T. Hamasaki, S.R. Evans, and K. Asakura. Design, data monitoring, and analysis of clinical trials with co-primary endpoints: A review. Journal of Biopharmaceutical Statistics, 28(1):28–51, 2018.

[105] S.K. Wang and A.A. Tsiatis. Approximately optimal one-parameter boundaries for group sequential trials. Biometrics, 43(1):193–199, 2016.

[106] S.J. Pocock. Group sequential methods in the design and analysis of clinical trials. Biometrika, 64(2):191–199, 1977.

[107] P.C. O'Brien and T.R. Fleming. A multiple testing procedure for clinical trials. Biometrics, pages 549–556, 1979.

[108] E.L. Lehmann and J.P. Romano. Generalizations of the familywise error rate. In Selected Works of EL Lehmann, pages 719–735. Springer, 2012.

[109] A. Dmitrienko, A.C. Tamhane, and F. Bretz. Multiple testing problems in pharmaceutical statistics. CRC press, 2009.

[110] P.F. Thall, R. Simon, and S.S. Ellenberg. A two-stage design for choosing among several experimental treatments and a control in clinical trials. Biometrics, 45(2):537–547, 1989.

[111] J.M.S. Wason and T. Jaki. Optimal design of multi-arm multi-stage trials. Statistics in Medicine, 31(30):4269–4279, 2012.

[112] B.C. Kahan, S. Rehal, and S. Cro. Risk of selection bias in randomised trials. Trials, 16(1), 2015.

[113] K.F. Schulz, D.G. Altman, and D. Moher. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ, 340, 2010.

[114] C.R. Mehta and S.J. Pocock. Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statistics in Medicine, 30(28):3267–3284, Dec 2011.

[115] R. Lin and J.J. Lee. Novel Bayesian Adaptive Designs and Their Applications in Cancer Clinical Trials, pages 395–426. Springer International Publishing, Cham, 2020.

[116] V.E. Johnson and J.D. Cook. Bayesian design of single-arm phase II clinical trials with continuous monitoring. Clinical Trials, 6(3):217–226, 2009.

[117] P. Dutton, S.B. Love, L. Billingham, and A.B. Hassan. Analysis of phase II methodologies for single-arm clinical trials with multiple endpoints in rare cancers: An example in Ewing's sarcoma. Statistical Methods in Medical Research, 27(5):1451–1463, 2018.

[118] J.J. Lee and D.D. Liu. A predictive probability design for phase II cancer clinical trials. Clinical Trials, 5(2):93–106, 2008.

[119] J.J. Lee and C.T. Chu. Bayesian clinical trials in action. Statistics in Medicine, 31(25):2955–2972, 2012.

[120] C. Yap, L.J. Billingham, Y.K. Cheung, C. Craddock, and J. O'Quigley. Dose transition pathways: the missing link between complex dose-finding designs and simple decision-making. Clinical Cancer Research, 23(24):7440–7447, 2017.

[121] Positioning imatinib for pulmonary arterial hypertension (PIPAH). https://clinicaltrials.gov/ct2/show/NCT04416750, 2020.

[122] K.W. Prins and T. Thenappan. WHO group I pulmonary hypertension: Epidemiology and pathophysiology. Cardiology Clinics, 34(3):363, 2016.

[123] E. Bibbona and A. Rubba. Boundary crossing random walks, clinical trials, and multinomial sequential estimation. Sequential Analysis, 31(1):99–107, 2012.

[124] J.J. Hanfelt, R.S. Slack, and E.A. Gehan. A modification of Simon's optimal design for phase II trials when the criterion is median sample size. Controlled Clinical Trials, 20(6):555–566, 1999.

Appendix A

Further results

Expected loss by design type, scenarios 2 and 3

Heat maps of expected loss for the admissible designs of each design types are shown in Figures A.1 and A.2 for scenarios 2 and 3 respectively.

Admissible designs, by design type (scenarios 2 and 3)

For completeness, the range of admissible designs for each compared design for scenarios 2 and 3 is shown in Figures A.3 and A.4. Again, the overall results are similar across all three scenarios: the designs that employ SC generally contain more admissible designs than those that do not. For these designs, the admissible design regions often contain slopes parallel to the hypotenuse, suggesting that the admissible design may be more dependent on the weight of N than $ESS(p_0)$ or $ESS(p_1)$. In some cases, this is manifested in long, thin regions near the hypotenuse. Here, the admissible designs have the greatest maximum sample size of all the possible admissible designs, with maximum sample size decreasing as the weight of N increases (that is, in the bottom left corner), as could be expected. When the weight of N is not close to 1, the novel designs often have a maximum sample size similar to those of the Simon designs.

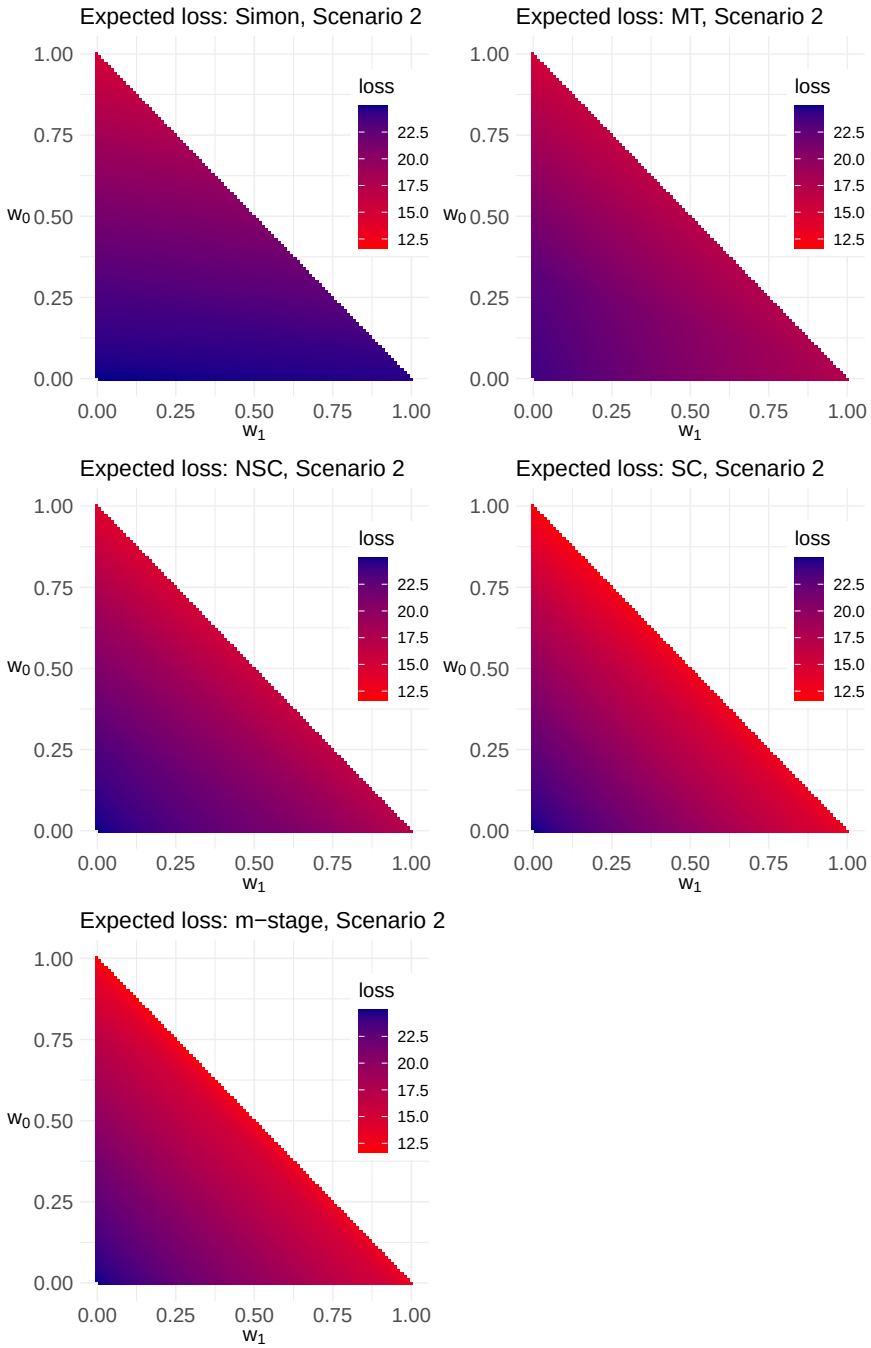


Fig. A.1 Expected loss for each design type, for scenario 2 $(\alpha, \beta, p_0, p_1) = (0.05, 0.20, 0.10, 0.30)$. MT: Mander and Thompson

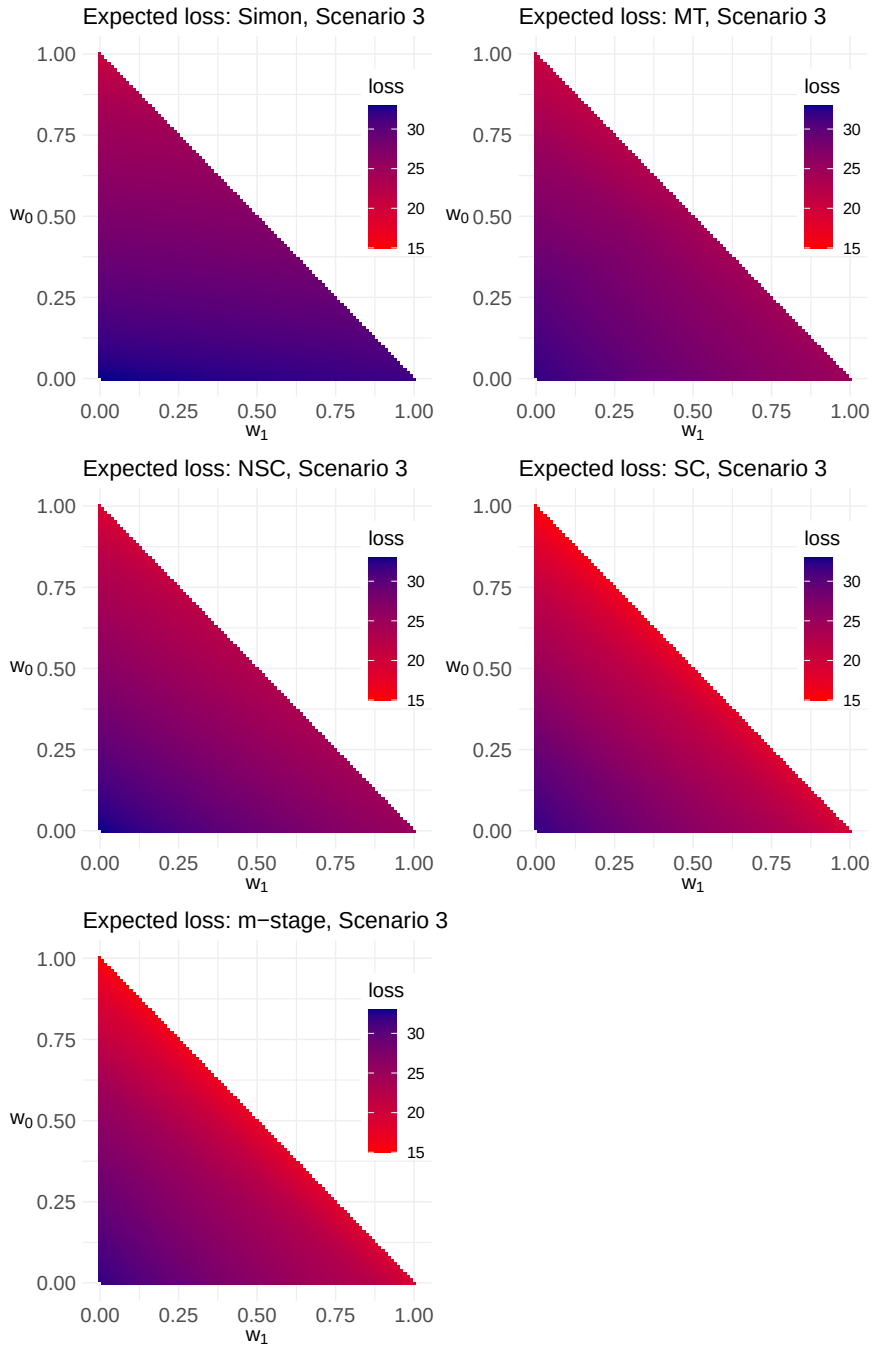


Fig. A.2 Expected loss for each design type, for scenario 3 $(\alpha, \beta, p_0, p_1) = (0.05, 0.20, 0.20, 0.40)$. MT: Mander and Thompson

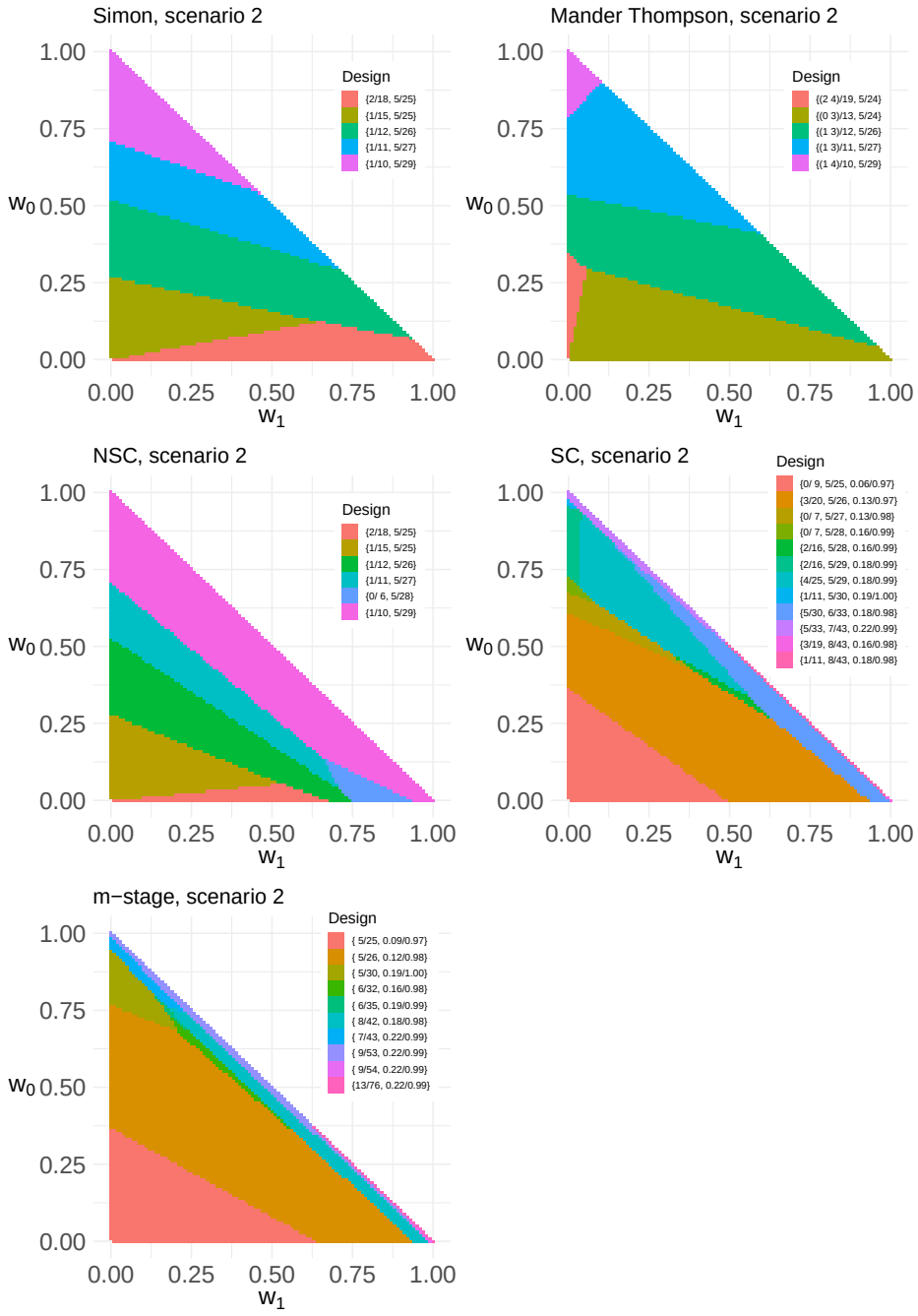


Fig. A.3 Admissible designs for each design type, for scenario 2 $(\alpha, \beta, p_0, p_1) = (0.05, 0.20, 0.10, 0.30)$.

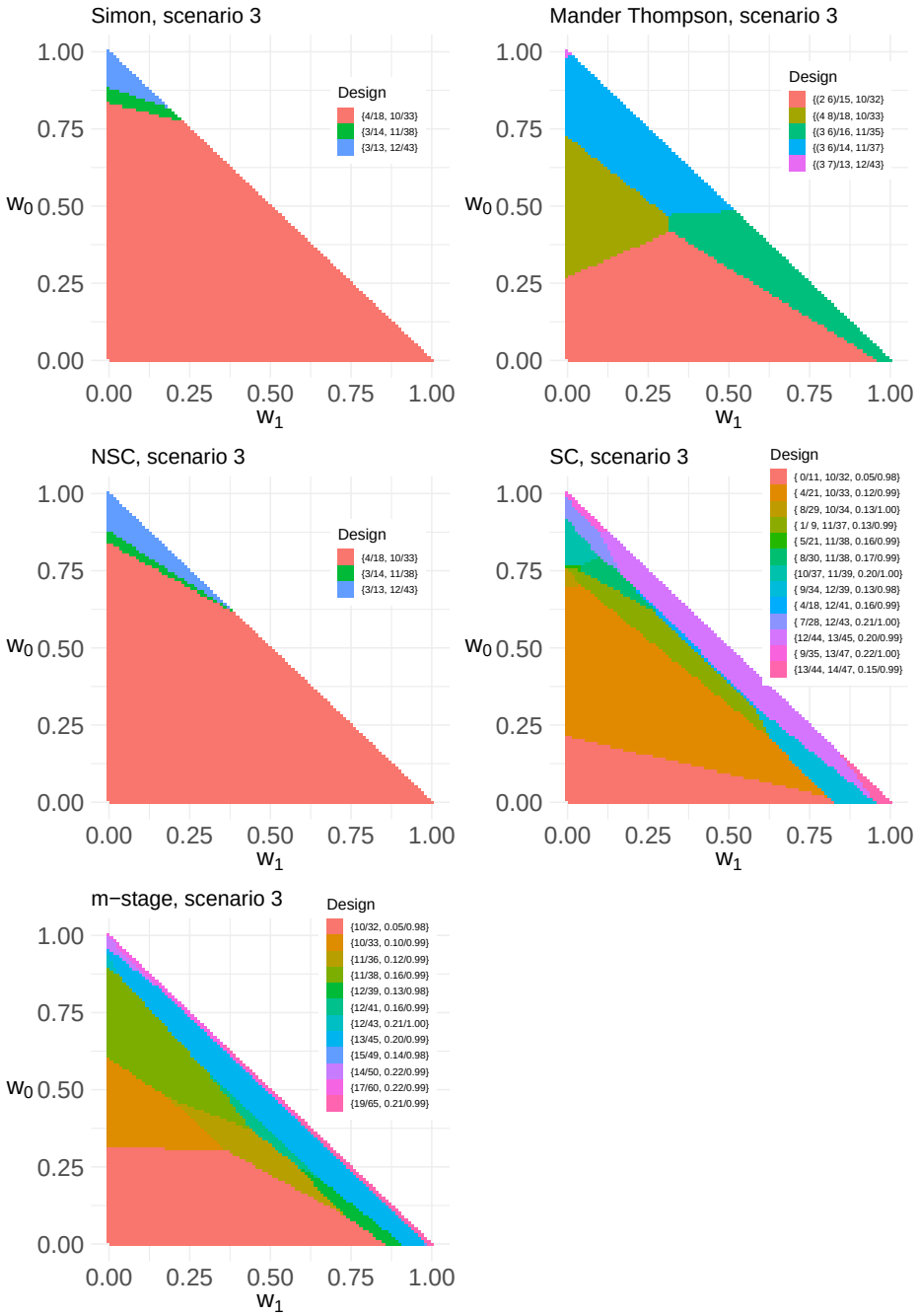


Fig. A.4 Admissible designs for each design type, for scenario 3 $(\alpha, \beta, p_0, p_1) = (0.05, 0.20, 0.20, 0.40)$.