University of Bremen

Doctoral Thesis

Bias and precision in early phase adaptive oncology studies and its consequences for confirmatory trials

Author:
Arsénio Quingue Nhacolo

Supervisor:
Prof. Dr. Werner Brannath

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Natural Sciences (Dr. rer. nat.)

in the

Competence Center for Clinical Trials Bremen
Faculty 03 (Mathematics and Computer Science)

August 30, 2018

Universität Bremen

Kompetenzzentrum für
Klinische Studien Bremen

Datum des Promotionskolloquiums: 8. Oktober 2018

Gutachter:
Prof. Dr. Werner Brannath (Universität Bremen, Deutschland)
Univ.-Prof. Mag. Dr. Martin Posch (Medizinische Universität Wien, Österreich)

UNIVERSITY OF BREMEN

Abstract

Faculty 03 (Mathematics and Computer Science)
Competence Center for Clinical Trials Bremen

Doctor of Natural Sciences (Dr. rer. nat.)

Bias and precision in early phase adaptive oncology studies and its consequences for confirmatory trials

by Arsénio Quingue NHACOLO

The need for a more efficient drug development process led to migration from the traditional fixed-sample clinical trial designs to group-sequential and adaptive designs, especially in early phases of clinical drug development. This, however, came with challenges in inference, since many of these newly proposed designs come without respective methods for statistical inference. In this dissertation, we study the estimation methods for oncology phase II group-sequential and adaptive designs in terms of bias and precision, and we propose new estimation methods for a new class of adaptive designs. We then evaluate the consequences, in terms of power, of using estimates from these designs to plan phase III trial. We also study and propose new approaches to adjust these estimates, based on the observed data, before employing them in planning of phase III sample size, in order to reach the desired power. Literature review showed that many estimation methods have been proposed for the classical single-arm two-stage group-sequential designs with a binary endpoint, which are the most commonly used designs in oncology trials of phase II. Simulation studies showed that the uniformly minimum variance unbiased estimator is the best amongst them in terms of bias and mean square error. However, for the adaptive group-sequential designs, these estimation methods have poor performance. Our proposed estimation methods in oncology phase II adaptive designs showed better performance as compared to the naïve maximum likelihood estimator. A direct use of estimates from phase II adaptive designs to plan phase III results in underpowered phase III trials. Therefore, adjusting (discounting) these estimates beforehand is necessary. The amount of discounting, however, depends on the estimator, with our proposed estimators requiring less discounting as compared to the naïve maximum likelihood estimator. Our proposed adjustment approaches show power improvements, which are similar across different estimators and design scenarios.

UNIVERSITÄT BREMEN

Zusammenfassung

Fachbereich 3 (Mathematik und Informatik)
Kompetenzzentrum für Klinische Studien Bremen

Doktor der Naturwissenschaften (Dr.rer.nat.)

Bias and precision in early phase adaptive oncology studies and its consequences for confirmatory trials

von Arsénio Quingue NHACOLO

Die Notwendigkeit eines effizienteren Medikamentenentwicklungsprozesses führte, insbesondere in frühen Phasen der klinischen Arzneimittelentwicklung, zu einer Abkehr von traditionellen Studiendesigns mit fester Fallzahl hin zu adaptiven gruppensequentiellen Designs. Jedoch gibt es für viele dieser neu vorgeschlagenen Studiendesigns noch keine geeigneten Methoden der statistischen Inferenz. In der vorliegenden Dissertation werden die Schätzmethoden für klassische gruppensequentielle Phase II Studien in der Onkologie in Bezug auf Verzerrung (Bias) und Präzision untersucht. Darüber hinaus wird die Eignung dieser Methodik, sowie die einer in der Dissertation neu hergeleiteten Schätzmethodik, für die Auswertung der adaptiven Versionen dieser Designs geprüft. Anschließend wird der Einfluss der Verwendung der aus diesen Studien erhaltenen Schätzern auf die Planung von Phase III Studien in Bezug auf die tatsächlich erzielte Power untersucht. Es existieren bereits Ansätze die Schätzer aus den Phase II Studien anzupassen um bei deren Verwendung in der Planung von Phase III Studien die gewünschte Power zu erreichen. Diese bereits bestehende und in dieser Dissertation neu hergeleitete Ansätze untersucht. Eine Literaturrecherche hat gezeigt, dass viele Schätzmethoden für die klassischen einarmigen zweistufigen Gruppensequentiellen Designs mit binären Endpunkten, welche die am häufigsten genutzten Designs in onkologischen Phase II Studien sind, vorgeschlagen werden. In der vorliegenden Dissertation werden Simulationsstudien durchgeführt die zeigen, dass der "uniformly minimum variance unbiased estimator" in Bezug auf Verzerrung und mittleren quadratischen Fehler der beste betrachtete Schätzer ist. Die Simulationsergebnisse zeigen jedoch, dass diese Schätzverfahren angewendet auf adaptive gruppensequentielle Designs zu schlechteren Ergebnissen führen. Die in dieser Dissertation vorgeschlagenen Schätzmethoden für adaptiven Phase II Studien in der Onkologie zeigten im Vergleich zum naiven Maximum-Likelihood-Schätzer bessere Ergebnisse. Eine direkte Verwendung der Schätzer aus Phase II Studien zur Planung von Phase III Studien führte zu einer geringeren als der geplanten Power. Daher ist eine Anpassung (Verkleinerung) dieser Schätzungen erforderlich. Die notwendige

Höhe der Anpassung hängt dabei von der Schätzmethodik ab, wobei der in dieser Dissertation vorgeschlagene Schätzer eine geringere Anpassung benötigt, als der naive Maximum-Likelihood-Schätzer. Unsere vorgeschlagenen Anpassungsansätze zeigen Leistungsverbesserungen, die sich über verschiedene Schätzer und Designszenarien hinweg ähneln.

Acknowledgements

I would like to thank my supervisor, Prof. Dr. Werner Brannath, for his attention and guidance during this project.

I would also like to thank Univ.-Prof. Mag. Dr. Martin Posch for being the second reviewer of this thesis.

My Doctoral Programme was possible thanks to the funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant.

Contents

List of Figures

List of Tables

Abbreviations

CAD	**C**onstant **A**rcsine **D**ifference
CDF	**C**umulative **D**istribution **F**unction
CI	**C**onfidence **I**nterval
CR	**C**omplete **R**esponse
CSSE	**C**onservative **S**ample **S**ize **E**stimation
DTD	**D**ual **T**hreshold **D**esign
FDA	**F**ood and **D**rug **A**dministration
FWER	**F**amily-**W**ise **E**rror **R**ate
GSD	**G**roup-**s**equential **D**esign
KIT	**KIT** Proto-Oncogene Receptor Tyrosine Kinase
MLE	**M**aximum **L**ikelihood **E**stimate
MUE	**M**edian **U**nbiased **E**stimator
ORR	**O**bjective **R**esponse **R**ate
PET	**P**robability of **E**arly **T**ermination
PP	**P**redictive **P**robability
PR	**P**artial **R**esponse
RCT	**R**andomized **C**linical **T**rial
RECIST	**R**esponse **E**valuation **C**riteria **I**n **S**olid **T**umours
RMSE	**R**oot **M**ean **S**quare **E**rror
ROC	**R**eceiver **O**perating **C**haracteristic
SA	**S**imulated **A**nnealing
SE	**S**tandard **E**rror
SC	**S**tochastically **C**urtailed
SCPRT	**S**equential **C**onditional **P**robability **R**atio **T**est
SP	**S**uccess **P**robability

STD	**S**ingle **T**hreshold **D**esign
TR	**T**otal **R**esponse
TTE	**T**ime **T**o **E**vent
UMVCUE	**U**niformly **M**inimum **V**ariance **C**onditionally **U**nbiased **E**stimator
UMVUE	**U**niformly **M**inimum **V**ariance **U**nbiased **E**stimator

To the loving memory of my father, Quingue Nhacolo, and to my mother, Celeste Nhancule, who despite many difficulties did their best to provide education to their children.

Chapter 1

Introduction

This chapter gives a brief motivational background and outlines the main scientific contributions of this dissertation. It gives, in addition, the details of how the rest of the dissertation is structured.

1.1 Motivation

Clinical drug development is a lengthy and costly process done in different phases (I, II, III and IV). Traditionally clinical trials were conducted using fixed-sample designs, in which all aspects of the trial are pre-specified and data analysis is only done at the end. The need to increase efficiency led to new design paradigms. For instance, in oncology phase II, the desire to expose less patients to futile treatments and accelerate the development of efficacious ones led to group-sequential designs. Group-sequential designs introduce interim analyses offering the possibility of early stopping for futility and/or efficacy (e.g., Shuster, 2002; Simon, 1989). Later, adaptive elements were added to these designs, allowing for modification of certain design aspects in the light of interim results (e.g., Englert and Kieser, 2013). However, the gains in efficiency comes at the cost of complicated inference. The traditional fixed-sample inference methods are not suitable for adaptive designs, and often yield treatment effect estimates that are biased and imprecise. Unfortunately most of the adaptive designs that are being proposed in the literature are mainly concerned with hypothesis testing and don't offer corresponding methods for estimating the efficacy parameter, leaving the implementers with no choice other than using the traditional fixed-sample methods. Bias and imprecision of estimates from phase II trials may result in poor planning of subsequent phase III trials, contributing to high failure rate in phase III, which is acknowledged to be as high as 50% in general (Pretorius, 2016) and 60% in oncology (Gan et al., 2012). Therefore, adequate estimation methods for phase II adaptive designs are important, the lack of which requires that caution and corrective measures be exercised when planning the corresponding phase III trials.

1.2 Contributions

The goals of this dissertation are to study phase II adaptive designs and methods for efficacy estimation in such designs, and to investigate the consequences of bias and imprecision of phase II efficacy estimates on the planning of phase III trials. New estimation methods for a class of adaptive phase II designs are proposed, and approaches for adjusting phase III sample size estimates are discussed. The main contribution of this dissertation are the following:

1. We studied the performance of different estimation methods proposed in the literature for oncology phase II group-sequential and adaptive designs. This was done via simulation studies, and included the commonly used single-arm designs with binary endpoint.
2. We proposed new estimation methods for oncology phase II adaptive designs. This was for two-stage adaptive designs with binary endpoint in which the second stage sample size and decision rules are functions of the number of successes in the first stage.
3. We analysed the consequences of using effect estimates from phase II adaptive design, which are often biased and imprecise, on planning phase III sample size and we discussed and proposed adjustment approaches in order to obtain adequately powered phase III trials.

1.3 Structure

The remainder of this dissertation is composed by four chapters, appendix and bibliography section. Chapter 2 contains a summarized literature review on phase II clinical trials designs, with an emphasis on oncology. In Chapter 3 we summarize efficacy estimation methods proposed in the literature for oncology phase II designs and we compare their performance via simulation studies. Then we propose new methods for interval and point estimation in adaptive designs, and compare them with the naïve maximum likelihood estimator (MLE). In Chapter 4 we turn into the issue of using effect estimates from phase II trials to plan later phase III trials. We first give a brief summary of what have been done in literature to deal with bias and imprecision of these estimates. Then we investigate the consequences, in terms of power, of using the estimates from adaptive phase II trials to plan phase III. This is done via simulations and include the estimates from the new methods we proposed in Chapter 3. Afterwards we propose a new approach to estimate the sample size adjustment factor to get a properly powered phase III trial accounting for the random results of phase II. Chapter 5 gives general summary and conclusion. Supplementary material is provided in the appendices.

Two papers and one `R` package were written as part of this dissertation. The `R` package documentation is in Appendix C and its corresponding package available on-line (`GitHub`). The first paper, which was published in the *Statistical Methods in Medical Research* journal, is in the Appendix D. The second paper is in the Appendix E, and it was still under submission for publication at the time of writing of this dissertation.

Chapter 2

Oncology phase II clinical trials designs

This chapter gives a literature review on oncology phase II clinical trials designs. First an overview of clinical trials phases is given along with summary of types of phase II designs. Then more details on designs commonly or likely to be used in oncology trials are given.

2.1 Overview

Drug development in humans comprises mainly three phases, phases I, II and III. **Phase I clinical trials** are the first in human studies, following preclinical development of a new therapeutic agent. The main goals in this phase is to assess tolerability and safety, and to find suitable dose(s) for the subsequent phases. Trials in this phase are usually conducted using healthy volunteers, except areas in which treatments are associated with severe side effects like oncology. Typically sample sizes are small. **Phase II clinical trials** have larger sample sizes, and are usually conducted using patients. This phase focus on the assessment of efficacy and safety, and it allows to make a decision of whether a therapy is worth further evaluation in later large scale phase III trials. **Phase III clinical trials** are also conducted using patients. They assess efficacy and safety, and aim at providing definitive evidence of treatment efficacy. This phase is primarily intended to support a licence submission to regulatory authorities. After a drug has received approval to be marketed, a phase IV trial may be conducted, which is usually a post-marketing surveillance trial primarily intended to detect rare or long-term adverse effects.

Apart from the development phases, clinical trials can also be grouped/classified using other characteristics. These include the number of treatment arms, statistical methods, number of interim analyses, adaptiveness, type and number of endpoints, etc.

Regarding the number of arms, trials can be of **single-arm** or **one-sample designs**, in which all patients receive the same treatment, with no control group and no variation in dose, formulation, or treatment regimen. These designs are common in oncology, where for ethical reasons a placebo control cannot be used, and patients may already have received unsuccessful treatments with standard therapies (Biswas et al., 2008). **Comparative** or **two-arm designs** are generally randomized trials, adding a concurrent control group to a single experimental group. Another category is **screening** or **selection** or **multi-arm designs**, in which several experimental treatments are compared, and possibly also compared with a concurrent control treatment.

Methodologically, clinical trials designs can be frequentist (classical), Bayesian or decision-theoretic. **Frequentist designs** focus on hypothesis testing and control of false positive and false negative error rates. The efficacy of the experimental treatment is summarized using a parameter θ. This parameter is then fixed, e.g., θ_0 under the null hypothesis (H_0), and inference focuses on comparing the observed random data with the distribution that would be expected if H_0 were true. In contrast, **Bayesian designs** consider θ as a random variable, and inference focuses on what can be said about its distribution. Being a random variable, θ has some distribution even before any data are observed, the prior distribution $P(\theta)$, which is updated by observing data X to get the posterior distribution $P(\theta|X)$. We can then obtain the posterior expected value of θ, or the posterior probability that it exceeds some specific value such as θ_0. **Decision-theoretic designs** model the decision-making process with the goal of getting an optimal decision that maximizes the value of some specified utility function. The utility function expresses the preferences of the decision-maker and is also a function of θ. $P(\theta|X)$ is used to calculate a posterior expected utility associated with each possible action that can be taken, then the action with the largest posterior expected utility is chosen.

With respect to presence or absence of interim analyses, we have **single-stage** or **fixed-sample designs**, in which the analysis is performed after a pre-planned number of patients has been accrued (Mariani and Marubini, 1996). **Sequential designs** are designs in which analysis is performed after the outcome of each new patient becomes available. At each step, a test statistic derived from the accumulated data is compared with an upper and lower test boundaries, allowing for early stopping (Mariani and Marubini, 1996). These designs are more efficient than the single-stage ones, however, difficult to implement since they require continuous monitoring of study results. As a compromise between the simplicity of single-stage designs and the efficiency of sequential designs, **multi-stage** or **group-sequential designs** accrue patients in batches (Mariani and Marubini, 1996). After each batch, the accumulated data is analysed and a decision is made to terminate the trial and draw a conclusion in favour of or against H_0, or to continue to the next stage[1].

It has been a tradition in clinical research to pre-specify all aspects of the trial beforehand and keep them unchanged during the trial conduct. This was in part due to the regulatory pressure to safeguard the validity and the integrity of clinical trials. However, designs that allow for modifications of trial as it progresses are becoming more commonly used, especially in early phase trials. **Adaptive design** is defined as a design that allows adaptations to trial and/or statistical procedures of the trial after its initiation without undermining the validity and integrity of the trial (Chow et al., 2005). The term **flexible design** is interchangeably used.

Another defining characteristic of clinical trial designs is the type and the number of endpoints. In phase II, the most popular types of endpoints are binary, ordinal, continuous and time-to-event. Most designs are with a single endpoint. In oncology this is often the treatment response. There are designs that consider a composite endpoint (i.e., endpoint that is made of more than one variable), and in oncology it can be a bivariate endpoint combining response and toxicity (e.g., Bryant and Day, 1995; Conaway and Petroni, 1995, 1996), or response and early progression (e.g., Sun et al., 2009; Zee et al., 1999).

Recently a new class of designs, termed *master protocol*, has emerged. A master protocol trial design have one overarching protocol with an objective of evaluating multiple treatments for one disease or one treatment for multiple diseases in multiple substudies (Sridhara et al., 2015). Master protocols can further be subdivided in *basket* (or *bucket*), *umbrella* and *platform* (or *standing*) trials. Basket trials aim at studying a single targeted therapy in the context of multiple diseases or disease subtypes (Woodcock and LaVange, 2017). In oncology, this trial tests one treatment simultaneously in different cancer types (baskets) possessing the same target genetic mutation (e.g., Cunanan et al., 2017; Leblanc et al., 2009). In an umbrella trial, multiple targeted therapies are studied in the context of a single disease (Woodcock and LaVange, 2017). In cancer research, an umbrella trial focus on a single tumour type, within which various molecular profiles are targeted (e.g., Barroilhet and Matulonis, 2018; Kim et al., 2011). Platform trial studies multiple targeted therapies in the context of a single disease in a perpetual manner, with therapies allowed to enter or leave the platform on the basis of a decision algorithm (Woodcock and LaVange, 2017). Platform trial follows a randomized design

[1] Some group-sequential designs only allow early termination for futility, i.e., when the conclusion is in favour of H_0

with a common control arm and many different experimental arms that enter and exit the trial as futility or efficacy are demonstrated (e.g., Hobbs et al., 2018; Lin and Bunn, 2017).

The classification of clinical trial designs is vast, here we have only listed the categories that we judged to be important to understand the subsequent contents of this dissertation. Furthermore, as it might have been obvious, these classification schemes do not lead to mutually exclusive groups of designs, many categories often overlap.

The number of oncology phase II designs that have been and are being proposed is overwhelmingly high. We found over 100 scientific papers proposing new designs or modifications of existing ones (see Appendix A for the complete list), with the majority being single-arm designs. Single-arm two-stage designs with binary endpoints are the most commonly used in practice (Englert and Kieser, 2012b), especially the optimal design by Simon (1989). Apart from being one of the most popular designs, Simon (1989)'s optimal design has inspired other designs including recent adaptive designs (e.g., Englert and Kieser, 2013; Shan et al., 2016a) that offer some efficiency improvements and have potential to be adopted in practice. Therefore, we study in detail these designs.

2.2 Single-arm group-sequential designs

Being a compromise between the simplicity of single-stage designs and the efficiency of sequential designs, group sequential designs are probably the most commonly used in phase II trials. The performance, in terms of sample size gains, of these designs increases with increasing number of stages. However, high number of stages results in high logistical complexity and administrative burden. In addition, the largest performance improvement is observed when moving from single-stage to two-stage designs (Rao et al., 2007), and this might be one of the reasons for the popularity of two-stage designs.

In general, oncology phase II single-arm group-sequential designs with binary endpoints test, at specific type I error rate α and type II error rate β, the null hypothesis (H_0) against the alternative (H_1),

$$H_0 : \pi \leq \pi_0 \text{ versus } H_1 : \pi \geq \pi_1 \tag{2.1}$$

where $\pi_1 > \pi_0$ and π is the response probability. π_0 is the maximum response rate considered clinically uninteresting and π_1 the minimum response rate considered to be of clinical interest. These designs are specified by the maximum number of stages, K, the number of subjects examined at each stage, (n_1, n_2, ..., n_K), the set of acceptance points, (l_1, l_2, ..., l_K), and the corresponding set of rejection points, (u_1, u_2, ..., u_K), with $u_k > l_k$ and $l_K = u_K - 1$, $k = 1, \ldots, K$. At stage k, the test statistic is $s_k = \sum_{i=1}^{k} x_i$, where x_i be the number of responses at stage i, then the procedure is to stop the trial and accept H_0 if $s_k \leq l_k$, or reject H_0 if $s_k \geq u_k$; otherwise continue to stage $k+1$. At the final stage H_0 is rejected if $s_K \geq u_K$ or, equivalently, if $s_K > l_K$ (Schultz et al., 1973).

Some authors (e.g., Chow and Chang, 2008; Mahajan and Gupta, 2010) label all group-sequential designs as adaptive designs. However, here we follow the approach by Bauer and Brannath (2004) and make distinction between the classical (non-adaptive) and adaptive group-sequential designs. The distinction lies essentially on the fact that, in classical group-sequential designs, interim analyses serve only the purpose of deciding whether to stop or continue the trial, while adaptive designs in addition offer the possibility of modifying trial characteristics based on interim results.

2.2.1 Classical two-stage designs with binary endpoints

As mentioned above, one of the most popular designs in oncology phase II are the designs by Simon (1989). They are single-arm designs testing the same hypotheses as in (2.1). Because of practical considerations in the management of multi-institution or multi-centre clinical trials, Simon (1989) restricted his attention to two-stage designs. He considered early stopping only for futility, arguing that when the drug has substantial activity ($\pi \geq \pi_1$) there is often interest in studying additional patients in order to estimate the proportion, extent, and durability of response. He proposed *optimal two-stage design*, which is a design that, given the parameters π_0 and π_1, satisfies the error probability constraints α and β, and minimizes the expected sample size (average sample number, ASN) when the response probability is π_0. Let n be the total sample size, n_1 the first stage sample size and n_2 the second stage sample size ($n_2 = n - n_2$). The expected sample size is calculated as

$$\text{ASN} = n_1 + (1 - \text{PET})n_2,$$

where PET is the probability of early termination after the first stage, calculated as $\text{PET} = B(l_1|\pi, n_1)$, where B denotes the cumulative binomial distribution, and π the true probability of response. The trial is terminated at the end of the first stage and the drug is rejected (H_0 is accepted) if l_1 or fewer patients out of n_1 respond to the treatment. The drug is rejected at the end of second stage if l or fewer responses out of total patients (n) are observed. Hence the probability of rejecting a drug with success probability π is

$$B(l_1|\pi, n_1) + \sum_{x=l_1+1}^{min[n_1,l]} b(x|\pi, n_1)B(l - x|\pi, n_2),$$

where b denotes the binomial probability mass function. The optimal designs are determined by enumeration using exact binomial probabilities. In addition, Simon (1989) determined also the *minimax two-stage design*, which minimizes the maximum sample size $n = n_1 + n_2$. For the same constraints (π_0, π_1, α, and β), the *minimax* design may be more preferable than the *optimal* design when the difference (between the two designs) in expected sample sizes is small and the patient accrual rate is low. Some examples of the designs are shown in Table 2.1.

Table 2.1: Simon (1989)'s designs for $(\alpha, \beta) = (0.05, 0.1)$

		Optimal Design				Minimax Design			
π_0	π_1	n_1	l_1	n	l	n_1	l_1	n	l
0.2	0.4	19	4	54	15	24	5	45	13
0.4	0.6	25	11	66	32	29	12	54	27

There are other group-sequential designs that preceded Simon (1989)'s designs. These include, among others, the pioneering work of Gehan (1961). In Gehan (1961)'s design, n_1 patients are accrued and treated in stage 1. If no response is observed, the trial is stopped and the treatment is discarded. Otherwise, the trial proceeds to stage 2, in which additional n_2 patients are accrued with the aim of estimating the response rate π with a desired precision.

Many other designs followed, some of which were inspired by Simon (1989)'s designs (see Table A.1 for more references).

2.2.2 Adaptive two-stage designs with binary endpoints

The classical phase II group-sequential designs as described above have fixed sample size and decision boundaries at each stage. Allowing the sample size and the corresponding decision boundaries of subsequent stages to depend on the results of previous stages may result in designs with improved efficiency. Such adaptive designs have been proposed by various authors (see Table A.1). Among them are the optimal adaptive two-stage designs proposed by Englert and Kieser (2013), testing the same hypotheses (2.1) and using the same optimality criteria (minimum expected sample size under the null hypothesis) as the Simon (1989)'s optimal design. These designs are defined by the first stage sample size, n_1, futility and efficacy boundaries, l_1 and u_1 ($u_1 > l_1$), which are fixed, and the second stage sample size, $n_2(x_1)$, which depends on the number of responses observed in the first stage, x_1. We further have the conditional error function, $D(x_1)$, and the corresponding decision boundary, $l(x_1)$, which are also functions of x_1. The final (second) stage efficacy boundary $u(x_1)$ is set to $u(x_1) = l(x_1) + 1$, with $l(x_1)$ being the futility boundary. $D(x_1)$ defines for each possible number of responses in the first stage, $x_1 \in \{0, ..., n_1\}$, the conditional type I error rate to be used in the second stage (Englert and Kieser, 2012b). At the interim analysis, the trial is stopped with failure to reject H_0 if $x_1 \leq l_1$ or with rejection of H_0 if $x_1 \geq u_1$. Otherwise the trial proceeds to the second (final) stage, after which H_0 is rejected if $p_2 \leq D(x_1)$ or, equivalently, $x > l(x_1)$, where p_2 is the second stage p-value, calculated with the $n_2(x_1)$ patients recruited after the first stage, and x is the total number of responses (i.e., x is the sum of x_1 and the number of responses observed in the second stage, x_2). Note that $x > l(x_1)$ is equivalent to $x \geq u(x_1)$. The discrete conditional error function $D(x_1)$ in Table 2.2 is non-decreasing in x_1, and takes values within $[0, 1]$.

The designs are found by exhaustive numerical search, fixing the response probability under the null and alternative hypotheses (π_0 and π_1), and the type I and type II error rates (α and β). An example of such designs is given in Table 2.2.

Table 2.2: Englert and Kieser (2013)'s optimal adaptive design for $(\pi_0, \pi_1, \alpha, \beta) = (0.2, 0.4, 0.05, 0.1)$

$n_1 = 20, n_{2,max} = 39$

x_1	$n_2(x_1)$	$D(x_1)$	$l(x_1)$
≤ 4	0	0	0
5	16	0.082	10
6	30	0.129	14
7	33	0.200	15
8	39	0.241	17
9	39	0.376	17
≥ 10	0	1	0

It is clear that in these designs the first stage decision boundaries are $l_1 = \max\{x_1 | D(x_1) = 0\} = \min\{x_1 | D(x_1) > 0\} - 1$ and $u_1 = \min\{x_1 | D(x_1) = 1\} = \max\{x_1 | D(x_1) < 1\} + 1$, and the first and second stage p-values are respectively $p_1 = 1 - B(x_1 - 1, n_1, \pi_0)$ and $p_2 = 1 - B(x_2 - 1, n_2(x_1), \pi_0)$, where $B(x, n, \pi)$ is the binomial cumulative distribution function (CDF) with x successes, n trials and success probability π .

Compared to the classical designs like Simon (1989)'s optimal design, these designs show improvements in terms of conservatism, being less conservative by better exhaustion of type I error and, therefore, showing savings on average sample size, which are attributed to the use of the discrete conditional error function and an improved search strategy.

Another recent proposal was by Shan et al. (2016a), who also proposed optimal adaptive two-stage designs similar to those of Englert and Kieser (2013). The differentiating characteristic of these designs is that the second stage's sample size is always a non-increasing function of the first stage's number of responses.

In these designs (Englert and Kieser, 2013; Shan et al., 2016a), represented by $\{n_1, n_2(x_1), l(x_1)\}$, the corresponding discrete conditional error function can be defined as $D(x_1) = 1 - B\left\{l(x_1) - x_1, \pi_0, n_2(x_1)\right\}$, $x_1 \in \{0, \ldots, n_1\}$. The overall type I error rate is given by

$$\sum_{x_1=0}^{n_1} D(x_1) \mathrm{Pr}_{H_0}(X_1 = x_1),$$

the type II error rate given by

$$1 - \sum_{x_1=0}^{n_1} \mathrm{Pr}_{H_1}\left\{p_{2,x_1} \leq D(x_1)\right\} \mathrm{Pr}_{H_1}(X_1 = x_1)$$

and the average sample size under the null hypothesis by

$$\mathrm{ASN}(\pi_0) = \sum_{x_1=0}^{n_1} \left\{n_1 + n_2(x_1)\right\} \mathrm{Pr}_{H_0}(X_1 = x_1).$$

The decision making based on the discrete conditional error function is identical to the one based on boundaries formulated in terms of the observed number of responses, therefore, the discrete conditional error function representation is sufficient to construct these designs. For any specified n_2, the set of the possible values of $D(x_1)$ is given by

$$\mathbb{P}_{2,n_2} := \{1 - B(x_2 - 1, \pi_0, n_2) | x_2 \in \{0, \ldots, n_2\}\}$$

and, for a given range of second-stage sample sizes N_2, the possible values of the discrete conditional error function are $\mathbb{P}_2 = \cup_{n_2 \in N_2} \mathbb{P}_{2,n_2} \cup \{0, 1\}$, where 0 and 1 represent the early stopping for futility or efficacy, respectively.

In general, designs are found by, for a suitable grid of n_1 and n_2, constructing a non-decreasing sequence $D(x_1) \in \mathbb{P}_2$, $x_1 \in \{0, \ldots, n_1\}$, and checking $D(x_1)$ for control of type I and type II error rates. The design minimizing the average sample size under H_0 among all designs satisfying the type I and II error constraints is selected.

Chapter 3

Estimation methods in oncology phase II designs

In this chapter we give an overview of methods for estimating efficacy parameter proposed in the literature for phase II oncology designs. We then compare, via simulations, the performance of some of these methods. Further we propose new estimation methods for a class of adaptive group-sequential designs, and study their performance using simulations. We have published part of this chapter as a research article in a peer-reviewed scientific journal.

3.1 Overview

The gains in efficiency and flexibility obtained by departing from the traditional fixed-sample clinical trials designs to group-sequential and adaptive designs come at a cost of complexity in inference methods. Many new proposals of such designs often come without the respective methods for estimation of the efficacy parameter. After implementing a trial following these designs, investigators are often left with no choice but performing inferential analysis using the traditional methods. By doing so, investigators risk jeopardizing the trial results, given that traditional methods may have poor performance when applied in a design context which they were not primarily intended for.

In phase II clinical trials, although the main goal is the hypothesis testing regarding the decision of whether a treatment is worth further investigation in a late large scale phase III trial, the estimation of efficacy remains important. This is especially so in cases where the treatment was deemed successfully since the efficacy estimate is needed for planning future trials. Most of the phase II trial designs in oncology are group-sequential, with some having in addition adaptive features (see Chapter 2). Due to the group-sequential nature (possibility of early stopping) and/or the adaptiveness of these designs, the naïve (fixed-sample) maximum likelihood estimate is no longer unbiased. This problem has been acknowledged by many authors, who proposed alternative estimates. Most of these alternative estimation methods are, however, design dependent. In the classical single-arm two-stage group-sequential designs with binary endpoint (discussed in detail in Chapter 2), different estimators have been proposed. These estimators include the bias-adjusted estimator (Chang et al., 1989) and its simplified version (Guo and Liu, 2005), the bias-corrected maximum likelihood estimator and the uniformly minimum variance unbiased estimator (Jung and Kim, 2004), the median unbiased estimator (Jovic and Whitehead, 2010; Koyama and Chen, 2008), the conditional maximum likelihood estimator (Tsai et al., 2008), the uniformly minimum variance conditionally unbiased estimator (Pepe et al., 2009), and the mean square error reduced estimator (Li, 2011). The conditional estimators of Tsai et al. (2008) and Pepe et al. (2009) are restricted to cases where trials proceeded to the second (final) stage. However, for the adaptive versions of these designs, little has been done regarding estimation. To our knowledge, it is only recently that some estimation approaches where discussed by Kunzmann and Kieser (2017).

Literature on estimation for group-sequential and adaptive designs common in areas other than oncology phase II is, however, vast. We give a brief summary of estimation methods proposed for various designs in Appendix B.

3.2 Simulation study

3.2.1 Introduction

It is clear from the overview above that there are constant research efforts aiming at providing inference methods that are adequate to group-sequential and adaptive design approaches. These methods are helpful tools for the investigators who want to benefit from the flexibility of such designs. However, as seen above, some of these designs end-up having quite a lot of alternative methods. With this multiplicity of methods, investigators are also faced with the problem of which method to use. This problem can be solved by doing extensive comparisons of the performance of the methods. Some authors (e.g., Bowden and Wason, 2012; Porcher and Desseaux, 2012) have taken step forward in this direction by conducting simulation studies.

Simon (1989)'s designs are very popular and widely used for oncology single-arm phase II clinical trials (Koyama and Chen, 2008). Owing to this popularity, various methods to estimate the efficacy parameter (response rate) after such designs have been proposed. In fact, the majority of the estimation methods for classical group-sequential designs listed in Section 3.1 are applicable to Simon (1989)'s designs. We conduct a simulation study to evaluate these methods. The main objectives are to compare the performance of estimation methods for Simon (1989)'s designs, and to study their behaviour when used for adaptive designs. For comparability of the results, we used adaptive designs that are similar to Simon (1989)'s, but with second stage sample size updated based on interim results using discrete conditional error function (Englert and Kieser, 2012b).

3.2.2 Methodology

3.2.2.1 Designs

A Simon (1989)'s design (described in detail in Section 2.2.1) testing the hypothesis about the response rate π, $H_0 : \pi \leq \pi_0$ versus $H_1 : \pi \geq \pi_1$, with specified type I and type II error rates, α and β, consists of first stage and overall sample sizes , n_1 and n, and their respective decision boundaries, l_1 and l. At the interim analysis, the trial is stopped with failure to reject H_0 if the number of responses x_1 is at most l_1, otherwise the trial continues to the last (second) stage. At the end, efficacy is concluded if the cumulative number of responses x is greater than l. The sample sizes and decisions boundaries are all pre-defined and fixed.

The adaptive design is obtained from the Simon (1989)'s design using the approach by Englert and Kieser (2012b). Let p_1 and p_2 be the first and second stage p-values, respectively. In analogy to the method proposed by Müller and Schäfer (2001, 2004), Englert and Kieser (2012b) defined for two-stage designs with discrete outcomes the discrete conditional error function D as a non-increasing function $D(p) : [0, 1] \to [0, 1]$ with support $\boldsymbol{P}_1$ and

$$\sum_{p \in \boldsymbol{P}_1} D(p) \mathrm{Pr}_{H_0}(P_1 = p) \leq \alpha$$

where $\boldsymbol{P}_1$ denotes the finite set of possible outcomes p_1 of the corresponding random variable P_1. For designs with binary endpoint, the p-values of the two stages with x_1 and x_2 observed responses in the first and second stage, respectively, are given by

$$p_1(x_1) = \mathrm{Pr}_{H_0}(X_1 \geq x_1) = 1 - B(x_1 - 1, \pi_0, n_1),$$

and

$$p_2(x_2) = \mathrm{Pr}_{H_0}(X_2 \geq x_2) = 1 - B(x_2 - 1, \pi_0, n_2),$$

where X_i denotes the random variable of the number of responses in stage i, $i = 1, 2$, and B the cumulative binomial distribution. The actual α level, α', of such a design is given by

$$\alpha' = \sum_{x_1=0}^{n_1} \text{CE}(x_1).\text{Pr}_{H_0}\{P_1 = p_1(x_1)\},$$

where

$$\text{CE}(x_1) = \begin{cases} 0 & \text{if } x_1 \leq l_1 \\ 1 - \{B(l - x_1, \pi_0, n_2)\} & \text{if } l_1 < x_1 < u_1 \\ 1 & \text{if } x_1 \geq u_1 \end{cases}$$

defines the conditional type I error rate when x_1 responses are observed in the first stage. In Simon (1989)'s designs, since there is no early stopping for efficacy, u_1 is set to any number greater than n_1, e.g., $u_1 = n_1 + 1$.

For any two-stage design with $\alpha' \leq \alpha$ a discrete conditional error function can be defined by $D(p_1(x_1)) = \text{CE}(x_1)$. By using this conditional error function, arbitrary design modifications after the first stage can be performed while still controlling the type I error rate. Therefore, to get adaptive versions of Simon (1989)'s designs, we apply the conditional error function approach to the original designs, and after the first stage we recalculate the sample size so that the conditional power, given the number of responses at the first stage, is equal to the target power $(1 - \beta)$.

3.2.2.2 Estimation methods

One of the reasons the fixed-sample maximum likelihood inference is not adequate to Simon (1989)'s designs and other similar designs is that apart from the number of responses, the stopping stage also plays a role in the outcome of the trial. This also defines the sample space configuration of the outcome, impacting the way probabilities are calculated. Let M and S be the random variables of the stopping stage m and the total number of successes (responses) s, respectively, where $s = x_1$ if $m = 1$ and $s = x = x_1 + x_2$ if $m = 2$. Let n be the total sample size ($n = n_1$ if $m = 1$), and $n_2 = n - n_1$ the second stage sample size. Jung and Kim (2004) have demonstrated that (M, S) is a complete and sufficient statistic for π, and that for Simon (1989)'s designs the probability mass function of (M, S) is

$$f_\pi(m, s) = \begin{cases} \pi^s(1-\pi)^{n_1-s}\binom{n_1}{s} & \text{if } m = 1 \\ \pi^s(1-\pi)^{n-s}\sum_{x_1=max(l_1+1,s-n_2)}^{min(s,n_1)}\binom{n_1}{x_1}\binom{n_2}{s-x_1} & \text{if } m = 2 \end{cases}$$

The naïve (fixed-sample) maximum likelihood estimator (MLE), which ignores the group-sequential nature of the designs, is the sample proportion which is given by $\hat{\pi}_{ml} = s/n$. As mentioned earlier, this estimator is biased. Using multiple summation Jennison and Turnbull (1983) as well as Chang et al. (1989) determined the bias of this estimator for

K-stage group-sequential designs to be

$$\text{bias}(\pi) = E(\hat{\pi}_{ml}|\pi) - \pi = \left[\sum_{m=1}^{K} \left(\sum_{x=l_{m-1}+1}^{l_m} + \sum_{x=u_m}^{u_{m-1}+n_m} \right) x f_\pi(m,x)/N_m \right] - \pi,$$

where N_m is the cumulative sample size at stage m, l_m and u_m are the futility and efficacy boundaries at stage m, with $l_0 = -1$ and $u_0 = 0$. In Simon (1989)'s designs where the number of stages is $K = 2$ and there is no early stopping for efficacy (i.e., $u_1 > n_1$), with $l_2 = l$ and $u_2 = u = l + 1$, the bias expression becomes

$$\text{bias}(\pi) = \frac{1}{n_1} \sum_{x=0}^{l_1} x f_\pi(1,x) + \frac{1}{n} \sum_{x=l_1+1}^{n} x f_\pi(2,x) - \pi.$$

Making use of this bias expression, two estimators with reduced bias were proposed. The first estimator, $\hat{\pi}_{ba}$, was proposed by Chang et al. (1989) and is defined as

$$\hat{\pi}_{ba} = \hat{\pi}_{ml} - \text{bias}(\hat{\pi}_{ba}).$$

This estimator is difficult to compute since the bias is evaluated at the same quantity that is being estimated. For this reason, Guo and Liu (2005) proposed another estimator, $\hat{\pi}_{br}$, that can be easily computed by evaluating the bias at the fixed-sample MLE:

$$\hat{\pi}_{br} = \hat{\pi}_{ml} - b(\hat{\pi}_{ml}).$$

Because of this computational simplicity, we consider only the estimator $\hat{\pi}_{br}$ for simulations.

Using RaoBlackwell theorem, Jung and Kim (2004) proposed the uniformly minimum variance unbiased estimator (UMVUE). Since the first stage sample proportion $\hat{\pi}_{ml1} = x_1/n_1$ is an unbiased estimator of π, the UMVUE ($\hat{\pi}_{umvu}$) of π is obtained as the conditional expectation of $\hat{\pi}_{ml1}$ given the sufficient statistic (M, S), i.e., $\hat{\pi}_{umvu} = E\left\{\hat{\pi}_{ml1}|(m,s)\right\}$. For Simon (1989)'s designs this estimator is explicitly defined as

$$\hat{\pi}_{umvu} = \begin{cases} \frac{s}{n_1} & \text{if } m = 1 \\ \frac{\sum_{x_1=max(l_1+1,s-n_2)}^{min(s,n_1)} \binom{n_1-1}{x_1-1}\binom{n_2}{s-x_1}}{\sum_{x_1=max(l_1+1,s-n_2)}^{min(s,n_1)} \binom{n_1}{x_1}\binom{n_2}{s-x_1}} & \text{if } m = 2 \end{cases}$$

Noting that the sample proportion based only on second stage data $\hat{\pi}_{ml2} = x_2/n_2$ is unbiased and that, conditional on $m = 2$, the overall sample proportion $\hat{\pi}_{ml} = s/n$ is a complete and sufficient statistic for the distribution of $\hat{\pi}_{ml2}$, Pepe et al. (2009) proposed the uniformly minimum variance conditionally unbiased estimator (UMVCUE) also using RaoBlackwell theorem. UMVCUE is meant for estimation conditional on continuing to the second stage. It is defined as $\hat{\pi}_{umvcu} = E\left\{\hat{\pi}_{ml2}|(2,s)\right\}$. For Simon (1989)'s designs its explicit expression is

$$\hat{\pi}_{umvcu} = \frac{\sum_{x_1=max(l_1+1,s-n_2)}^{min(s,n_1)} \binom{n_1}{x_1}\binom{n_2-1}{s-x_1-1}}{\sum_{x_1=max(l_1+1,s-n_2)}^{min(s,n_1)} \binom{n_1}{x_1}\binom{n_2}{s-x_1}}.$$

In simulations scenarios for unconditional estimation (i.e., estimation done irrespective of whether the trial stopped at first stage or continued to second stage), we set $\hat{\pi}_{umvcu} = \frac{s}{n_1}$ if $m = 1$ to allow comparison with other estimators.

Koyama and Chen (2008) proposed to get the point estimate of π for Simon (1989)'s designs using the overall p-value (based on the data from both stages). First they noted that it is a common practice to incorrectly compute a p-value at the end of second stage assuming the data were collected in a single stage. They call this the conventional p-value, and it is expressed as

$$p_c = \sum_{x_1=0}^{n_1} \mathrm{Pr}_{\pi_0}\left[X_1 = x_1\right]\mathrm{Pr}_{\pi_0}\left[X_2 \geq s - x_1\right].$$

It is clear that this p-value is incorrect. This is seen by the fact that the summand includes impossible sample paths in which $X_1 < l_1$ and $X_2 = s - X_1$. They proposed a preferred p-value which does not include these impossible sample paths, expressed as

$$p_p = \begin{cases} \mathrm{Pr}_{\pi_0}(X_1 \geq s) & \text{if } m = 1 \\ \sum_{x_1=l_1+1}^{n_1} \mathrm{Pr}_{\pi_0}(X_1 = x_1)\mathrm{Pr}_{\pi_0}(X_1 \geq s - x_1) & \text{if } m = 2 \end{cases}$$

This p-value is consistent with stage-wise ordering (we elaborate more on stage-wise ordering later in Section 3.3.2).

Then they proposed a median unbiased estimator, denoted $\hat{\pi}_{mu}$, which is the value of the response rate under H_0 ($\tilde{\pi}_0$) that, given the observed data, would give a p-value of 0.5, i.e.,

$$\hat{\pi}_{mu} = \left\{\tilde{\pi}_0 : p_p(\tilde{\pi}_0) = 0.5\right\},$$

where $p_p(\tilde{\pi}_0)$ means that π_0 is replaced by $\tilde{\pi}_0$ in the expression of p_p.

3.2.2.3 Simulation set-up

The performance of the estimator was in terms of bias and root mean square error (RMSE). To cover some spectrum of Simon (1989)'s designs, both optimal and minimax (see Section 2.2.1), 9 scenarios were considered per design: the response rate under the null hypothesis $\pi_0 = 0.1, 0.2, 0.3$, with the difference under the alternative $\pi_1 - \pi_0 = 0.1, 0.2, 0.3$ for each π_0. The assumed type I and type II error rates were $\alpha = 0.05$ and $\beta = 0.1$. To check the effect of departures from the assumed π_1, trials assuming lower and higher (than π_1) true response rate π where also simulated for each design scenario.

The adaptive versions of the designs (see Section 3.2.2.1) where obtained based on the classical ones (as defined above) by, at the interim analysis, updating the second stage sample size so as to achieve a conditional power of 90%. The final decision rule was based on conditional type I error, defined using the original design parameters (Englert and Kieser, 2012b).

Unconditional and conditional estimation was performed with all estimators. By unconditional estimation we mean that the efficacy parameter π is estimated irrespective of whether the trial stopped at first stage or continued to the second stage, and conditional estimation means that estimation is done only if the trial proceeded to the second stage. To serve as benchmark for comparison, the sample proportion based on first stage data only, $\hat{\pi}_{ml_1}$, and based on second stage data only, $\hat{\pi}_{ml_2}$, were also computed, respectively, for unconditional and conditional estimation. For each scenario, 10000 replicate trials were generated.

3.2.3 Results

Table 3.1 shows the performance of estimators for a Simon (1989)'s optimal design, with true response rate (π) equal to that under H_1 (π_1). The results of the corresponding adaptive version of the design (obtained as explained in Section 3.2.2.3) are shown in Table 3.2. Figures 3.1 and 3.2 show, respectively, the unconditional and conditional performance for the same optimal design for different values of π.

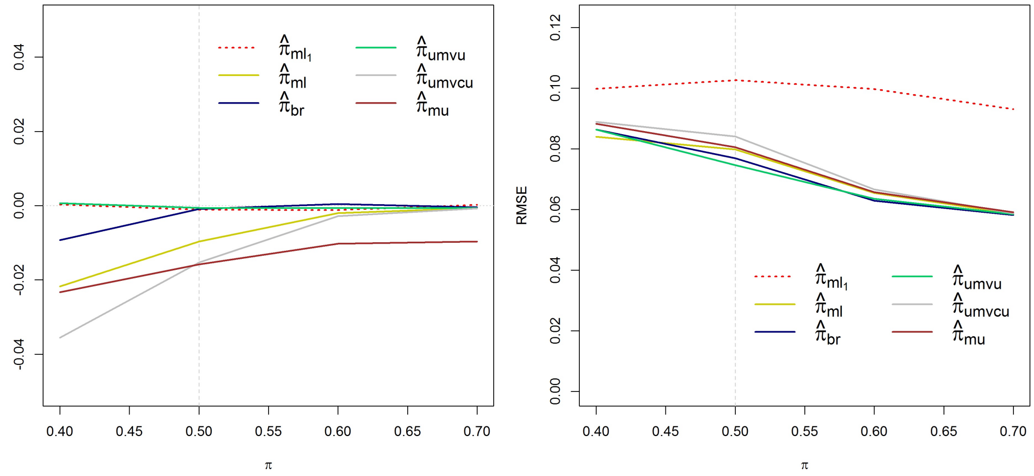
Table 3.1: Performance of estimators for Simon (1989)'s optimal design with parameters $\pi_0 = 0.3$, $\pi_1 = 0.5$, $\pi = 0.5$, $\alpha = 0.05$, $\beta = 0.1$. The "Unconditional" column holds the performance of estimators for estimation done irrespective of the stopping stage, while "Conditional" is for estimation done only for the trials that stopped at second (final) stage.

	Unconditional		Conditional	
Estimator	**Bias**	**RMSE**	**Bias**	**RMSE**
$\hat{\pi}_{ml_i}$*	-0.00114	0.10260	-0.00002	0.08036
$\hat{\pi}_{ml}$	-0.00968	0.07977	0.00568	0.06048
$\hat{\pi}_{br}$	-0.00089	0.07683	0.01521	0.05620
$\hat{\pi}_{umvu}$	-0.00058	0.07457	0.01555	0.05279
$\hat{\pi}_{umvcu}$	-0.01529	0.08410	-0.00039	0.06652
$\hat{\pi}_{mu}$	-0.01583	0.08050	0.00052	0.05643

*$i = 1$ for unconditional and $i = 2$ for conditional estimation

For unconditional estimation in classical Simon's designs, the UMVUE ($\hat{\pi}_{umvu}$) outperformed all other estimators in terms of bias. It was followed by the bias-reduced estimator ($\hat{\pi}_{br}$), which had negligible bias. Regarding the RMSE, $\hat{\pi}_{umvu}$ was again the best performer, closely followed by $\hat{\pi}_{br}$. Conditional on moving to the second stage, the UMVCUE ($\hat{\pi}_{umvcu}$) was the winner in terms of bias, but worse in terms of RMSE (although still better than $\hat{\pi}_{ml_2}$). Here the $\hat{\pi}_{umvu}$ and $\hat{\pi}_{br}$ were the worse in terms of bias, but better than $\hat{\pi}_{umvcu}$ in terms of RMSE.

In general, bias is negative for unconditional estimation and positive for conditional. The bias moves and stabilizes towards zero as π moves from π_0 to values greater than π_1. The exception is for $\hat{\pi}_{mu}$, having bias crossing the zero line.



(a) Bias

(b) RMSE

FIGURE 3.1: Performance of estimators (unconditional) for different values of true response rate π. Trials were simulated under Simon (1989)'s optimal design with parameters $\pi_0 = 0.3$, $\pi_1 = 0.5$, $\alpha = 0.05$, $\beta = 0.1$. Estimation was done in all the simulated trials

For the adaptive versions of the designs there was no clear winner. For π closer to π_1, $\hat{\pi}_{umvcu}$ tended to have lower bias, but in most cases it was worse than $\hat{\pi}_{ml_1}$. Regarding the RMSE, the median unbiased estimator ($\hat{\pi}_{mu}$), $\hat{\pi}_{umvu}$ and $\hat{\pi}_{br}$ were better than others but no clear winner amongst them. In conditional estimation the bias was even higher, with $\hat{\pi}_{mu}$ tending to have lower bias for π near π_1, followed by $\hat{\pi}_{umvcu}$. $\hat{\pi}_{mu}$ had also the lowest RMSE, followed by $\hat{\pi}_{umvu}$, $\hat{\pi}_{br}$ and $\hat{\pi}_{ml}$.

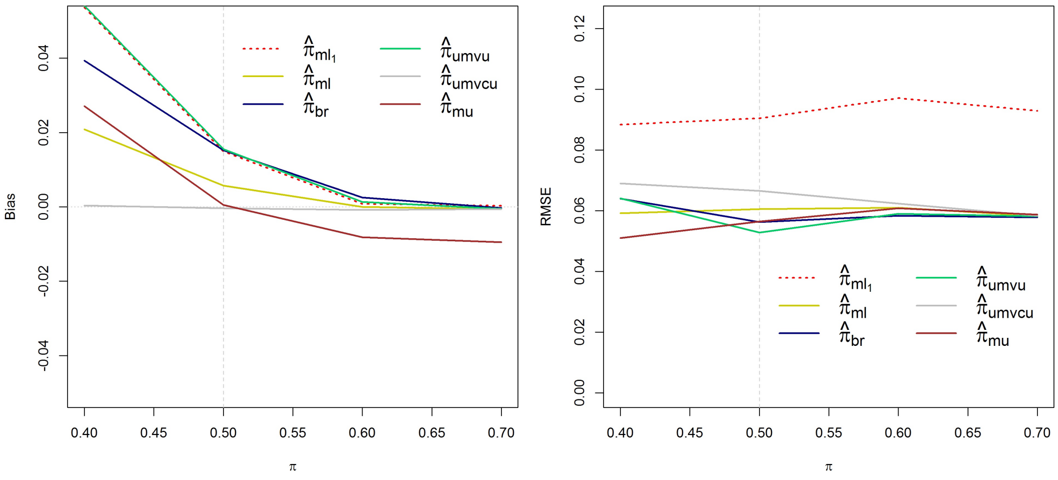
3.2.4 Conclusion

For the classical Simon's designs, the UMVUE is recommended for unconditional estimation and, the UMVCUE for conditional estimation, as they outperform the other estimators.

TABLE 3.2: Performance of estimators for adaptive version of Simon (1989)'s optimal design with parameters $p_0 = 0.3$, $p_1 = 0.5$, $p = 0.5$, $\alpha = 0.05$, $\beta = 0.1$. The "Unconditional" column holds the performance of estimators for estimation done irrespective of the stopping stage, while "Conditional" is for estimation done only for the trials that stopped at second (final) stage.

	Unconditional		Conditional	
Estimator	**Bias**	**RMSE**	**Bias**	**RMSE**
$\hat{\pi}_{ml_i}*$	0.00051	0.10215	-0.00110	0.15142
$\hat{\pi}_{ml}$	0.00644	0.09491	0.02327	0.08074
$\hat{\pi}_{br}$	0.01256	0.09258	0.02990	0.07776
$\hat{\pi}_{umvu}$	0.01266	0.09094	0.03001	0.07562
$\hat{\pi}_{umvcu}$	0.00094	0.09940	0.01731	0.08640
$\hat{\pi}_{mu}$	-0.00474	0.09156	0.01265	0.07237

$^*i = 1$ for unconditional and $i = 2$ for conditional estimation



(a) Bias

(b) RMSE

FIGURE 3.2: Performance of estimators (conditional on proceeding to second stage) for different values of true response rate π. Trials were simulated under Simon (1989)'s optimal design with parameters $\pi_0 = 0.3$, $\pi_1 = 0.5$, $\alpha = 0.05$, $\beta = 0.1$

For the adaptive versions of Simon's designs, there is no clearly best estimator to recommend. This is due to the fact that none of the estimators was developed for this type of designs. In fact, these estimators are not theoretically valid for such designs since they were derived with the assumption that both the first and second stage sample sizes are pre-defied and fixed. However in the adaptive designs considered, the second stage sample size is recalculated given the results of the first stage. Even if simulations had shown good performance of some of them, we would not recommend them since there is no sound theoretical reasoning to support their validity. Therefore, more research to find adequate estimation methods is needed.

3.3 New estimation methods for adaptive designs

As seen in the previous sections, the estimation methods proposed for oncology phase II group-sequential designs (GSD) are not applicable to their adaptive counterparts. Many estimation methods for adaptive GSD have been proposed in the literature (e.g., Bebu et al., 2013, 2010; Bowden et al., 2014; Bowden and Glimm, 2008, 2014; Bowden and Trippa, 2015; Brannath et al., 2006; Broberg and Miller, 2017; Carreras and Brannath, 2013; Cheng and Shen, 2004; Coburger and Wassmer, 2003; Gao et al., 2013; Kimani et al., 2015; Luo et al., 2012; Posch et al., 2005; Stallard and Todd, 2005). However, most of these methods are intended for phase III clinical trials designs, and little has been done in phase II. To the best of our knowledge, the discussion on estimation for the Simon-like phase II adaptive designs was only done recently by Kunzmann and Kieser (2017). They investigated different strategies and proposed a method for estimating the response rate. Their proposed point estimator uses the Bayesian framework and it can be interpreted as a constrained posterior mean estimate based on the non-informative Jeffreys prior.

As an alternative to the Bayesian procedure by Kunzmann and Kieser (2017), we propose a frequentist procedure for interval and point estimation. This procedure is for single-arm adaptive GSD with binary endpoint, in which the sample size of the second stage is a pre-defined function of the number of responses in the first stage. Some of the approaches we propose, however, can be extended to designs with flexible adaptation rules. We use the concept of stage-wise ordering to defined our procedure. Therefore, first we propose and discuss different sample space orderings approaches, from which we derive methods for calculating an overall p-value and then the interval and point estimates. In the following subsections, we provide a brief summary of the targeted designs, then we elaborate on our proposed methods. This is followed by the results of a simulation study, and we close the section with conclusions and a discussion.

We have published the methods discussed here in the *Statistical Methods in Medical Research* journal (see Nhacolo and Brannath, 2018, and the Appendix D)

3.3.1 Design characteristics

We construct and discuss our methods for adaptive designs similar to those proposed by Englert and Kieser (2013) and Shan et al. (2016a). These designs are described in Section 2.2.2, for convenience we summarize them here. They are single-arm two-stage designs with a binary endpoint, and are intended to be used for oncology phase II trials. Unlike the classical GSD, they allow the second stage sample size to vary with the number of the observed responses in the first stage. The hypotheses tested in these designs are

$$H_0 : \pi \leq \pi_0 \text{ versus } H_1 : \pi \geq \pi_1,$$

where H_0 and H_1 are the null and the alternative hypothesis, respectively, π_0 is the maximum response rate considered to be uninteresting and π_1 is the minimum desirable response rate ($\pi_1 > \pi_0$).
The designs are built to satisfy specific type I and type II error rates constraints (α and β), and are defined by fixed and varying elements. The fixed elements are the first stage sample size, n_1, futility and efficacy boundaries, l_1 and u_1 ($u_1 > l_1$). The second stage contains the varying elements, namely the second stage sample size, $n_2(x_1)$, the conditional error function, $D(x_1)$ and the corresponding decision boundary, $l(x_1)$, which are pre-specified functions of the number of responses observed in the first stage, x_1. $D(x_1)$ defines for each possible number of responses in the first stage, $x_1 \in \{0, ..., n_1\}$, the conditional type I error rate to be used in the second stage (Englert and Kieser, 2012b).

The trial is stopped at the first stage if $x_1 \leq l_1$ (H_0 not rejected) or if $x_1 \geq u_1$ (H_0 rejected); otherwise the trial proceeds to the second stage, after which H_0 is rejected if $p_2 \leq D(x_1)$ or, equivalently, $x > l(x_1)$, where p_2 is the second stage p-value and $x = x_1 + x_2$, with x_2 being the number of responses observed in the second stage. Note that the second stage rejection boundary is set to $u(x_1) = l(x_1) + 1$, therefore $x > l(x_1)$ is equivalent to $x \geq u(x_1)$. An example of such designs is given in Table 3.3.

TABLE 3.3: Englert and Kieser (2013)'s optimal adaptive design for $(\pi_0, \pi_1, \alpha, \beta) = (0.2, 0.4, 0.05, 0.1)$

$n_1 = 20, n_{2,max} = 39$			
x_1	$n_2(x_1)$	$D(x_1)$	$l(x_1)$
≤ 4	0	0	0
5	16	0.082	10
6	30	0.129	14
7	33	0.200	15
8	39	0.241	17
9	39	0.376	17
≥ 10	0	1	0

The discrete conditional error function $D(x_1)$ is non-decreasing in x_1, and takes values within $[0, 1]$. We assume these two properties in our methodology. As it can be seen from the design example in Table 3.3, the first stage decision boundaries are $l_1 = \max\{x_1 | D(x_1) = 0\}$ and $u_1 = \min\{x_1 | D(x_1) = 1\}$, and the first and second stage p-values are $p_1 = 1 - B(x_1 - 1, n_1, \pi_0)$ and $p_2 = 1 - B(x_2 - 1, n_2(x_1), \pi_0)$, where $B(x, n, \pi)$ is the binomial cumulative distribution function with x successes, n trials and success probability π.

3.3.2 Classical sample space orderings

A sample space ordering is necessary for the calculation of the probability of obtaining an outcome that is at least as extreme as the observed one, which is in turn fundamental for the construction of confidence intervals and p-values. For the traditional single-variable outcomes from fixed-sample designs, the sample space ordering is simply the ordering of the real numbers. However, in GSD the sample space ordering is not straightforward. This is because, apart from the test statistic, the stopping stage also plays a role when ordering the outcomes. Various sample space orderings have been proposed by different authors for GSD. These include the stage-wise ordering, first proposed by Armitage (1957) and later discussed by other authors (Fairbanks and Madsen, 1982; Jennison and Turnbull, 2000; Siegmund, 1978; Tsiatis et al., 1984; Wassmer and Brannath, 2016), the likelihood ratio ordering (Chang, 1989; Chang and O'Brien, 1986; Rosner and Tsiatis, 1988), the sample mean ordering (Emerson and Fleming, 1990), and the score test ordering (Rosner and Tsiatis, 1988).

The stage-wise ordering is a widely used ordering in GSD. For classical GSD counterparts of the adaptive designs above, i.e., designs in which n_2 and l are also fixed, this ordering can be defined as follows. Let m be the stopping stage and x the total number of responses. A trial outcome (m', x') is at least as extreme (against H_0) as the observed trial outcome (m, x), written as $(m', x') \succeq (m, x)$, if one of the following conditions is met:

$$\begin{aligned} &\text{(A)} \quad m' = m \text{ and } x' \geq x \\ &\text{(B)} \quad m' = 1,\, m = 2 \text{ and } x' \geq u_1 \\ &\text{(C)} \quad m' = 2,\, m = 1 \text{ and } x \leq l_1 \end{aligned}$$

For the adaptive designs, due to the nature of the conditional error function, the stage-wise ordering discussed here can be inconsistent with the design's decision rule when $m' = m = 2$. For example, for the design in Table 3.3, the minimum total number of responses necessary to reject H_0 is 11 if $x_1 = 5$ and 18 if $x_1 = 8$. If we have two outcomes, say an outcome $Y = (2, 13)$ with $x_1 = 5$ and another outcome $Y' = (2, 16)$ with $x_1 = 8$, according to the stage-wise ordering $Y' \succeq Y$, although Y leads to the rejection of H_0 and Y' doesn't.

3.3.3 Alternative sample space orderings

Alternative sample space orderings that solve the inconstancies discussed above are needed for the adaptive GSD. Here we propose a new sample space orderings that take into account the conditional error function and adaptation rule of the design. We define the orderings as follows. If both outcomes are from trials that stopped at the final (second) stage (i.e., $m' = m = 2$), we order them taking into account their respective rejection boundaries. To achieve this we define a function of the trial outcome, denoted $\delta(x_1, x_2)$, that in some way incorporates its respective rejection boundary. In all other cases we order the outcomes as in the classical stage-wise ordering discussed above (Section 3.3.2). Then it follows that $(m', x_1', x') \succeq (m, x_1, x)$ if one of the following conditions is satisfied:

(A1) $m' = m = 1$ and $x' \geq x$

(A2) $m' = m = 2$ and $\delta(x_1', x_2') \geq \delta(x_1, x_2)$

(B) $m' = 1$, $m = 2$ and $x' \geq u_1$

(C) $m' = 2$, $m = 1$ and $x \leq l_1$

Three different methods for defining $\delta(x_1, x_2)$ are proposed. The first two methods quantify the deviation between x and $l(x_1)$. In the first one $\delta(x_1, x_2)$ is defined using x directly as

$$\delta(x_1, x_2) = x_1 + x_2 - l(x_1) = x - l(x_1) \tag{3.1}$$

and in the second one $\delta(x_1, x_2)$ is defined using the second stage p-value as

$$\delta(x_1, x_2) = \tilde{\delta}\left[x_1, p_2(x_2)\right] = D(x_1) - p_2(x_2) \tag{3.2}$$

Both methods define δ such that it equals to a constant when the outcome is at the decision boundary (i.e., when $x_2 = l(x_1) - x_1$ and $p_2 = D(x_1)$). That is, $\delta\left[x_1, l(x_1) - x_1\right] = c_1$ and $\delta\left[x_1, D(x_1)\right] = c_2$. Here $c_1 = c_2 = 0$, meaning that the null hypothesis is rejected if $\delta(x_1, x_2) > 0$. The inequality $\delta(x_1', x_2') \geq \delta(x_1, x_2)$ in the case (A2) of our proposed sample space ordering can be stated as $x_2' \geq x - l(x_1) + l(x_1') - x_1'$ for the first method and $p_2' \leq p_2 - D(x_1) + D(x_1')$ for the second one.

The function δ as defined in (3.1) and (3.2) is strictly linked to the design's decision rules and, therefore, requires that the trial design be strictly followed.

Next we define δ using combination functions from adaptive tests to allow for flexibility. Combination functions combine the first and the second stage p-values, with the assumption that the data from the two stages are from independent cohorts of patients. An extensive discussion on adaptive combination tests can be found in Wassmer and Brannath (2016). We define a combination function $C(p_1, p_2)$, which is non-decreasing in both arguments and continuous in p_2, setting the early stopping boundaries $\alpha_0 = 1 - B(l_1 - 1, n_1, \pi_0)$ and $\alpha_1 = 1 - B(u_1 - 1, n_1, \pi_0)$, and finding a critical value c

such that the type I error is controlled, i.e.,

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 I_{[C(p_1,p_2)\leq c]} dp_2 dp_1 = \alpha$$

where $I_{[S]}$ equals to 1 if S is true and 0 otherwise.

Given the combination test, the most natural ordering on the second stage is according to $C(p_1, p_2)$, i.e., if we have performed a second stage we consider a second trial outcome with stage-wise p-values (p'_1, p'_2) as more extreme than our observed outcome if $C(p'_1, p'_2) < C(p_1, p_2)$. Even though the phase II designs we are dealing with might not be based on a combination function C we can build an ordering based on C that is consistent with the rejection region given by the function D (or equivalently given by the function l). We accomplish this by defining the corresponding conditional error function of C

$$A(p_1) = max\{y \in [0,1] : C(p_1, y) \leq c\}$$

and calculate the backward image p_{1_b} such that $A(p_{1_b}) = D(x_1)$, where $D(x_1)$ is the conditional error of the original design. Then we use p_{1_b} instead of p_1 in the combination function.

A natural and common choice for $C(p_1, p_2)$ is the weighted inverse normal combination function (Lehmacher and Wassmer, 1999), which can be represented as (Wassmer and Brannath, 2016)

$$C(p_1, p_2) = 1 - \Phi\left[w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2)\right],$$

where Φ is standard normal CDF, and w_1 and w_2 are predefined weights chosen such that $w_1^2 + w_2^2 = 1$. Here we propose to use weights that give more emphasis to the stage with higher sample size, i.e.,

$$w_1 = w_1(x_1) = \sqrt{\frac{n_1}{n_1 + n_2(x_1)}} \text{ and } w_2 = w_2(x_1) = \sqrt{\frac{n_2(x_1)}{n_1 + n_2(x_1)}}.$$

The conditional error function of the inverse normal combination function is

$$A(p_1) = 1 - \Phi\left[\frac{\Phi^{-1}(1-c) - w_1\Phi^{-1}(1-p_1)}{w_2}\right].$$

We solve the equation $A(p_{1b}) = D(x_1)$ for p_{1b}:

$$\begin{aligned}
&A(p_{1b}(x_1)) = D(x_1)\\
\Leftrightarrow &1 - \Phi\left\{\frac{\Phi^{-1}(1-c) - w_1(x_1)\Phi^{-1}\left[1 - p_{1b}(x_1)\right]}{w_2(x_1)}\right\} = D(x_1)\\
\Leftrightarrow &\frac{\Phi^{-1}(1-c) - w_1(x_1)\Phi^{-1}\left[1 - p_{1b}(x_1)\right]}{w_2(x_1)} = \Phi^{-1}\left[1 - D(x_1)\right]\\
\Leftrightarrow &\Phi^{-1}\left[1 - p_{1b}(x_1)\right] = \frac{\Phi^{-1}(1-c) - w_2(x_1)\Phi^{-1}\left[1 - D(x_1)\right]}{w_1(x_1)}\\
\Leftrightarrow &1 - p_{1b}(x_1) = \Phi\left\{\frac{\Phi^{-1}(1-c) - w_2(x_1)\Phi^{-1}\left[1 - D(x_1)\right]}{w_1(x_1)}\right\}\\
\Leftrightarrow &p_{1b}(x_1) = 1 - \Phi\left\{\frac{\Phi^{-1}(1-c) - w_2(x_1)\Phi^{-1}\left[1 - D(x_1)\right]}{w_1(x_1)}\right\}
\end{aligned}$$

We, finally, define δ as

$$\delta(x_1, x_2) = \bar{\delta}\left[p_{1b}(x_1), p_2(x_2)\right] = 1 - C(p_{1b}, p_2). \tag{3.3}$$

With δ defined in this way, the condition $\delta(x_1', x_2') \geq \delta(x_1, x_2)$ in the case (A2) of the proposed sample space ordering becomes $C(p_{1b}', p_2') \leq C(p_{1b}, p_2)$, meaning that the outcome with lower $C(p_{1b}, p_2)$ is considered to be more extreme. As it is the case with the other ordering discussed so far, this ordering is monotone in x_2, i.e., for two outcomes (x_1, x_2) and (x_1, x_2'), $x_2' > x_2$ implies that $\delta(x_1, x_2') > \delta(x_1, x_2)$. We have checked this property empirically for all the Englert and Kieser (2013)'s designs.

Ordering the trial outcomes by the proportion of responses, $(x_1+x_2)/[n_1+n_2(x_1)]$, might seem a plausible sample space ordering, however it would not guarantee consistency with the design's decision rule. This is because in some designs the ratio of the final decision boundary and the total sample size, $l(x_1)/[n_1 + n_2(x_1)]$, is not constant. For instance, in the Englert and Kieser (2013)'s design for $(\pi_0, \pi_1, \alpha, \beta) = (0.5, 0.7, 0.05, 0.2)$ this ratio is 0.6 if $x_1 = 14$ and 0.8 if $x_1 = 15$.

3.3.4 Overall p-value

Now that we have proposed new sample space orderings (Section 3.3.3), we use them to derive an overall p-value, denoted by Q, meant to be calculated when the trial has been terminated. First we introduce a modification of the outcome notation to explicitly include x_1, i.e., we change the notation from (m, x) to (m, x_1, x). This is because, as seen above, we need to explicitly know both x_1 and $x_2 = x - x_1$ to define δ, on which Q is based. We also introduce X_1 and X_2, the random variables for x_1 and x_2, respectively. Q is defined as the probability of observing under H_0 an outcome (m', x_1', x') that is similar or more extreme than the outcome (m, x_1, x) actually observed in the trial. If the observed outcome is from a trial that stopped at the first stage outcomes with $x_1' \geq x_1$ are more extreme, irrespective of their stopping stage, implying the overall p-value

$$Q = Q(x_1) = \Pr_{\pi_0}(X_1 \geq x_1).$$

If the observed outcome is from a trial that continued to the second stage, more extreme are outcomes from trials that stopped at the first stage with $x_1' \geq u_1$ or continued to the second stage with $\delta(x_1', x_2') \geq \delta(x_1, x_2)$, then

$$\begin{aligned} Q =& Q(x_1, x_2) = \\ & \Pr_{\pi_0}(X_1 \geq u_1) + \sum_{x_1'=l_1+1}^{u_1-1} \Pr_{\pi_0}(X_1 = x_1') \Pr_{\pi_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | X_1 = x_1'\right]. \end{aligned}$$

Since X_1 and X_2 follow binomial distribution, we can write the overall p-value as

$$Q = \begin{cases} 1 - B(x_1 - 1, n_1, \pi_0) & \text{if } m = 1 \\ 1 - B(u_1 - 1, n_1, \pi_0) + \sum\limits_{x_1'=l_1+1}^{u_1-1} b(x_1', n_1, \pi_0) \Pr_{\pi_0}\left(\Delta \geq \delta | x_1'\right) & \text{if } m = 2 \end{cases}$$

where $b(x, n, \pi)$ is the binomial probability mass function with x successes, n trials and success probability π, $\Delta = \delta(X_1, X_2)$ and $\delta = \delta(x_1, x_2)$.

We discuss in the following lines approaches to calculate the probability of $\delta(X_1, X_2) \geq \delta(x_1, x_2)$ under H_0, i.e., $\Pr_{\pi_0}[\delta(X_1, X_2) \geq \delta(x_1, x_2)]$.

Method 1:

For the $\delta(x_1, x_2)$ defined in (3.1), since we are working directly with the number of events (responses), this probability can easily be calculated as

$$\begin{aligned} &\Pr_{\pi_0}\left(\Delta \geq \delta | x_1'\right) \\ &\quad= \Pr_{\pi_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | x_1'\right] \\ &\quad= \Pr_{\pi_0}\left[X - l(X_1) \geq x - l(x_1) | x_1'\right] \\ &\quad= \Pr_{\pi_0}\left[X_1 + X_2 - l(X_1) \geq x - l(x_1) | x_1'\right] \\ &\quad= \Pr_{\pi_0}\left[X_2 \geq x - l(x_1) + l(X_1) - X_1 | x_1'\right] \\ &\quad= 1 - B\left[x - l(x_1) + l(x_1') - x_1' - 1, n_2(x_1'), \pi_0\right]. \end{aligned}$$

For the other two methods we use approximations. We make use of the fact that a p-value P is in general stochastically not smaller than a standard uniform variate, i.e.,

$$\Pr_{\pi_0}(P \leq \gamma) \leq \gamma,\ \gamma \in [0, 1].$$

Assuming that the first stage design is pre-fixed and is strictly followed, and that the first and the second stage data are from independent cohorts of patients, using the second stage p-value p_2 as the test statistic guarantees the *conditional invariance principle* (Wassmer and Brannath, 2016). This implies that conditional on the first stage data and second stage design, the distribution of p_2 under H_0 is not smaller than the uniform distribution. This, in turn, implies that the type I error rate is controlled irrespective of the adaptation rule.

Method 2:

When using the $\delta(x_1, x_2)$ in (3.2), we approximate $\Pr_{\pi_0}(\Delta \geq \delta)$ as

$$\begin{aligned}
Pr_{\pi_0}\left(\Delta \geq \delta | x_1'\right) =& \Pr_{\pi_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | x_1'\right] \\
=& \Pr_{\pi_0}\left[D(X_1) - P_2 \geq D(x_1) - p_2 | x_1'\right] \\
=& \Pr_{\pi_0}\left[P_2 \leq p_2 - D(x_1) + D(X_1) | x_1'\right] \\
\approx& \left\langle p_2 - D(x_1) + D(x_1')\right\rangle_{[0,1]}
\end{aligned}$$

where

$$\langle\omega\rangle_{[0,1]} = \begin{cases} 0 & \text{if } \omega < 0 \\ \omega & \text{if } 0 \leq \omega \leq 1 \\ 1 & \text{if } \omega > 1 \end{cases}$$

Method 2v2:

Another way of calculating $Pr_{\pi_0}(\Delta \geq \delta | x_1')$ in Method 2 is to use the fact that $p_2 = 1 - B\left[x_2 - 1, n_2(x_1), \pi_0\right]$ and the corresponding binomial quantile function, denoted by $B_q\left[p, n_2(x_1), \pi_0\right]$ (where p is a cumulative probability), as follows:

$$\begin{aligned}
Pr_{\pi_0}\left(\Delta \geq \delta | x_1'\right) =& \Pr_{\pi_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | x_1'\right] \\
=& \Pr_{\pi_0}\left[P_2 \leq p_2 - D(x_1) + D(X_1) | x_1'\right] \\
=& \Pr_{\pi_0}\left\{1 - B\left[X_2 - 1, n_2(X_1), \pi_0\right] \leq p_2 - D(x_1) + D(X_1) | \, x_1'\right\} \\
=& \Pr_{\pi_0}\left\{B\left[X_2 - 1, n_2(X_1), \pi_0\right] \geq 1 - p_2 + D(x_1) - D(X_1) | \, x_1'\right\} \\
=& \Pr_{\pi_0}\left\{X_2 - 1 \geq B_q\left[1 - p_2 + D(x_1) - D(X_1), n_2(X_1), \pi_0\right] | \, x_1'\right\} \\
=& \Pr_{\pi_0}\left\{X_2 \geq 1 + B_q\left[1 - p_2 + D(x_1) - D(X_1), n_2(X_1), \pi_0\right] | \, x_1'\right\} \\
=& 1 - B\left\{B_q\left[1 - p_2 + D(x_1) - D(x_1'), n_2(x_1'), \pi_0\right], n_2(x_1'), \pi_0\right\}
\end{aligned}$$

Method 3:

Finally, using the $\delta(x_1, x_2)$ in (3.3) we have that

$$\begin{aligned}\Pr_{\pi_0}\left(\Delta \geq \delta | x_1'\right) =& \Pr_{\pi_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | x_1'\right]\\ =& \Pr_{\pi_0}\left[C(P_{1b}, P_2) \leq C(p_{1b}, p_2) | x_1'\right]\\ =& \Pr_{\pi_0}\left[P_2 \leq 1 - \Phi(z_b)\right]\\ \approx& 1 - \Phi(z_b)\end{aligned}$$

where

$$z_b = \frac{w_1 \Phi^{-1}(1 - p_{1b}) + w_2 \Phi^{-1}(1 - p_2) - w_1' \Phi^{-1}(1 - p_{1b}')}{w_2'},$$

with $w_1 = w_1(x_1)$, $w_1' = w_1(x_1')$, $w_2 = w_2(x_1)$ and $w_2' = w_2(x_1')$. The expression of z_b is obtained by solving the inequation $\delta(x_1', x_2') \geq \delta(x_1, x_2)$:

$$\begin{aligned}&\delta(x_1', x_2') \geq \delta(x_1, x_2)\\ &\Leftrightarrow \delta\left[p_{1b}(x_1'), p_2(x_2')\right] \geq \delta\left[p_{1b}(x_1), p_2(x_2)\right]\\ &\Leftrightarrow 1 - C(p_{1b}', p_2') \geq 1 - C(p_{1b}, p_2)\\ &\Leftrightarrow C(p_{1b}', p_2') \leq C(p_{1b}, p_2)\\ &\Leftrightarrow 1 - \Phi\left[w_1' \Phi^{-1}\left(1 - p_{1b}'\right) + w_2' \Phi^{-1}\left(1 - p_2'\right)\right] \leq 1 - \Phi\left[w_1 \Phi^{-1}\left(1 - p_{1b}\right) + w_2 \Phi^{-1}\left(1 - p_2\right)\right]\\ &\Leftrightarrow \Phi\left[w_1' \Phi^{-1}\left(1 - p_{1b}'\right) + w_2' \Phi^{-1}\left(1 - p_2'\right)\right] \geq \Phi\left[w_1 \Phi^{-1}\left(1 - p_{1b}\right) + w_2 \Phi^{-1}\left(1 - p_2\right)\right]\\ &\Leftrightarrow w_1' \Phi^{-1}\left(1 - p_{1b}'\right) + w_2' \Phi^{-1}\left(1 - p_2'\right) \geq w_1 \Phi^{-1}\left(1 - p_{1b}\right) + w_2 \Phi^{-1}\left(1 - p_2\right)\\ &\Leftrightarrow \Phi^{-1}\left(1 - p_2'\right) \geq \frac{w_1 \Phi^{-1}\left(1 - p_{1b}\right) + w_2 \Phi^{-1}\left(1 - p_2\right) - w_1' \Phi^{-1}\left(1 - p_{1b}'\right)}{w_2'}\\ &\Leftrightarrow 1 - p_2' \geq \Phi\left[\frac{w_1 \Phi^{-1}\left(1 - p_{1b}\right) + w_2 \Phi^{-1}\left(1 - p_2\right) - w_1' \Phi^{-1}\left(1 - p_{1b}'\right)}{w_2'}\right]\\ &\Leftrightarrow p_2' \leq 1 - \Phi\left[\frac{w_1 \Phi^{-1}\left(1 - p_{1b}\right) + w_2 \Phi^{-1}\left(1 - p_2\right) - w_1' \Phi^{-1}\left(1 - p_{1b}'\right)}{w_2'}\right]\\ &\Leftrightarrow p_2' \leq 1 - \Phi(z_b)\end{aligned}$$

Note that some of steps above are only valid because Φ is a continuous and monotone increasing function.

3.3.5 Point and interval estimation

Here we used the overall p-value defined above to derive interval and point estimates. Following the the approach discussed in Wassmer and Brannath (2016, Chapter 8), we construct the confidence interval (CI) by considering all the null hypotheses

$$H_0^{\tilde{\pi}_0} : \pi \leq \tilde{\pi}_0, \text{ with } 0 \leq \tilde{\pi}_0 \leq 1.$$

The region $\{\tilde{\pi}_0 : Q(\tilde{\pi}_0) = \Pr_{\tilde{\pi}_0}[(M, X_1, X) \succeq (m, x_1, x)] > \alpha\}$ is a one-sided $(1-\alpha)100\%$ CI defined as $]\pi_L^{\alpha}; 1]$, where the lower bound π_L^{α} is the solution, in $\tilde{\pi}_0$, of the equation $Q(\tilde{\pi}_0) = \alpha$. That is, the CI is a set of $\tilde{\pi}_0$ for which H_0 is not rejected.

As the point estimate we take the lower bound of the 50% one-sided CI, i.e., $\hat{\pi} = \pi_L^{0.5}$, which is an approximate median unbiased estimator. Similar estimators have been proposed for classical oncology two-stage GSDs by Koyama and Chen (2008) and Jovic and Whitehead (2010), which are applicable only if n_2 is a constant for all x_1.

In order to this estimation technique to work it is necessary that the overall p-value $Q(\pi)$ as function of response probability π be monotone increasing for $\pi \in [0, 1]$. We checked the monotonicity of $Q(\pi)$ numerically for all 34 designs listed in Englert and Kieser (2013), for all possible outcomes and π ranging from 0 to 1 by increments of 0.01. We found that Methods 2 and 3 are monotone in all designs. Method 1 is monotone, except for four designs when $\pi \geq 0.8$. In the case of non-monotonicity, a conservative solution may be found using the cumulative maximum of $Q(\pi)$, i.e., $Q_{cm}(\pi) = max\{Q(\pi') : \pi' \leq \pi\}$.

3.3.6 Simulation study

We conducted an extensive simulation study to evaluate different aspects of our proposed methods. We checked the behaviour of the overall p-value (Q) for all possible outcomes in all designs listed in Englert and Kieser (2013), with varying value of the response rate under the null (from 0 to 1 by increments of 0.01). Then we evaluated the performance of the point estimates in terms of bias and root mean square error (RMSE), and the performance of the confidence intervals in terms of coverage probability and mean of the lower bound. We also estimated the type I error and power using the original decision rule and using the overall p-value from the proposed methods.

The bias was calculated as $\frac{1}{T}\sum_{t=1}^{T}(\hat{\pi}_t - \pi)$ and RMSE as $\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\pi}_t - \pi)^2}$, where T is the total number of simulated trials, $\hat{\pi}$ the estimated response probability and π the true response probability (under which trials were simulated). The coverage probability was computed as the proportion of trials in which the $(1-\alpha)100\%$ CI contained the true response rate π. The type I error was calculated as the proportion of trials simulated under $\pi = \pi_0$ in which H_0 was rejected, and the power calculated similarly but for trials simulated under $\pi = \pi_1$.

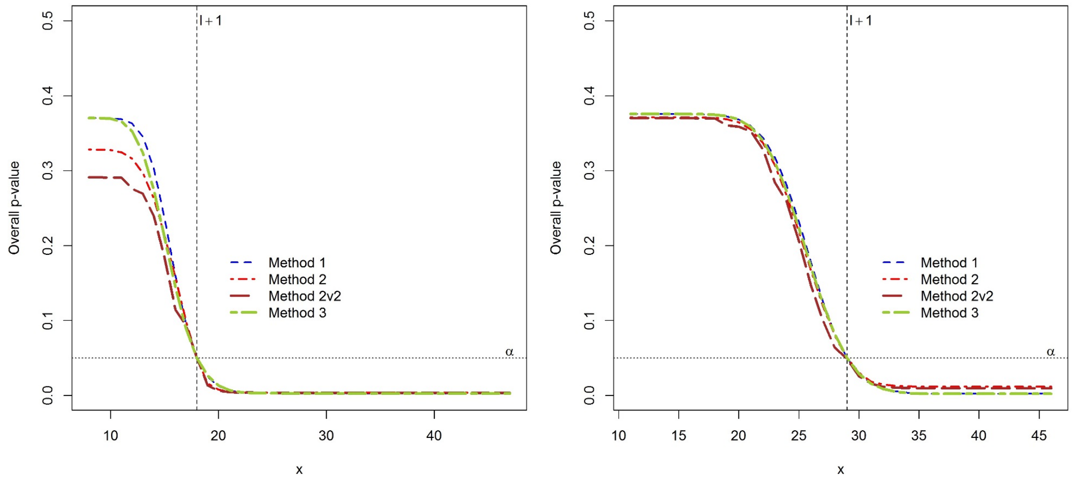
We have included, for comparison purposes, the naïve maximum likelihood estimator (MLE), which ignores the adaptiveness of the design (i.e., it assumes the data is from a fixed-sample design). This estimator is likely to be employed when analysing data from adaptive designs for which no specific estimation methods are available. We used two versions of the naïve MLE, one that uses all trial data, $\hat{\pi}_p = [x_1 + x_2]/[n_1 + n_2(x_1)]$, and the other that uses the first stage data only, $\hat{\pi}_{p1} = x_1/n_1$. The reason for including $\hat{\pi}_{p1}$ is that since it is unbiased, it will serve as benchmark for comparison with respect to RMSE, i.e., a new estimator would not be desirable if it would be outperformed by $\hat{\pi}_{p1}$ in terms of RMSE. We denote the estimated response probability by $\hat{\pi}_{m1}$ for Method 1, $\hat{\pi}_{m2}$ for Method 2, $\hat{\pi}_{m2v2}$ for Method 2v2, and $\hat{\pi}_{m3}$ for Method 3. The simulation were done for two designs of Englert and Kieser (2013), one with a moderate π_1, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, we call this design 1, and the other (design 2) with relatively high π_1, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. For both designs we varied, in the simulated trials, the true response probability π from 0 to 1 by increments of 0.01. For each scenario 50 000 trials were simulated.

We have implemented all the methods described above in the statistical programming language `R` (R Core Team, 2017), and included in a package described in Appendix C.

3.3.6.1 Results

Results on overall p-value showed that p-values from all the proposed methods are consistent with design's decision rule. However, due to approximations used for calculation of these p-values in some methods, there will be some borderline cases in which the overall p-value may lead to a different conclusion. Part of the p-value results, for all possible values of x when $x_1 = 8$ and $x_1 = 11$ in two designs, are shown in Figure 3.3. Results in Table 3.4 show that type I error rate and power of Methods 1 and 2v2 are equal to those of design's original decision rule, which are in turn close to the nominal levels. Methods 2 and 3 are conservative, as their type I error rate is lower compared to other methods.

The Figures 3.4 and 3.5 show, respectively, the results on bias and RMSE of the estimators for values of π ranging from 0 to 1. The Table 3.5 shows the results of simulations under H_1 ($\pi = \pi_1$), and, in addition to bias and RMSE, it shows the mean and the first, second and third quartiles of the estimates, and the coverage probability and the mean lower bound of the one-sided $(1 - \alpha)100\%$ CI. The Table 3.6 presents the same results for simulations under $\pi = \pi_1 + 0.1$. The point estimators behave differently depending on how close or far the true response rate (π) is from the value under the alternative hypothesis (π_1). For values of π close to π_0, all the estimators are negatively mean



(a) Design 1

(b) Design 2

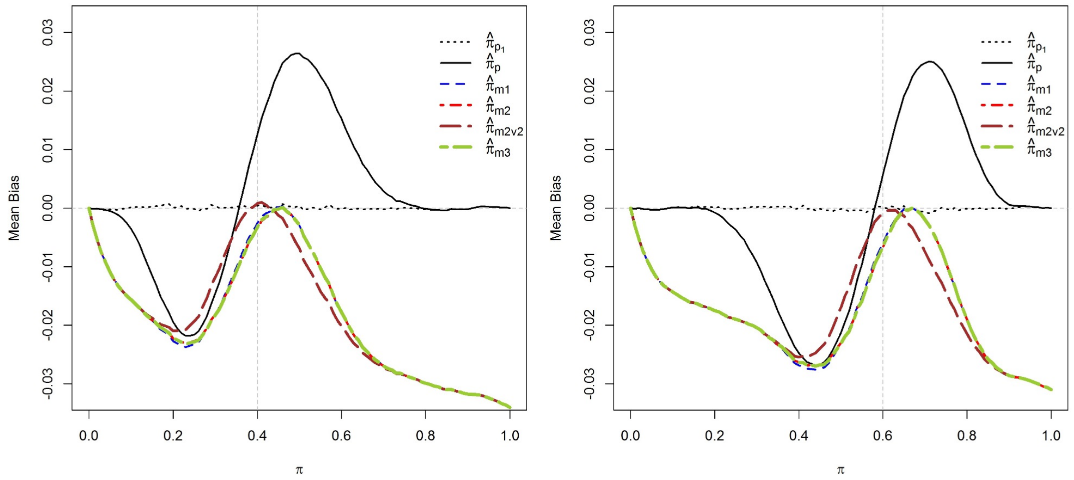
FIGURE 3.3: Plot of overall p-value (Q) as function of the total number of responses (x). Figure (a) is of the design $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$ with $x_1 = 5$, and (b) of $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$ with $x_1 = 10$. Both are cases where the trial continues to second stage. The vertical line represents the minimum total number of responses necessary to reject H_0 using design's decision rule.

TABLE 3.4: Type I error rate and power based on design's original decision rule (Orig.) and on the overall p-value from the proposed methods (Met.), from 50000 simulation runs. The two first rows are for design 1, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and the last two for the design 2, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$.

	Decision rule				
	Orig.	Met. 1	Met. 2	Met.2v2	Met. 3
Type I error	0.0503	0.0503	0.0430	0.0503	0.0430
Power	0.9002	0.9002	0.8877	0.9002	0.8877
Type I error	0.0492	0.0492	0.0421	0.0492	0.0428
Power	0.8990	0.8990	0.8884	0.8990	0.8892

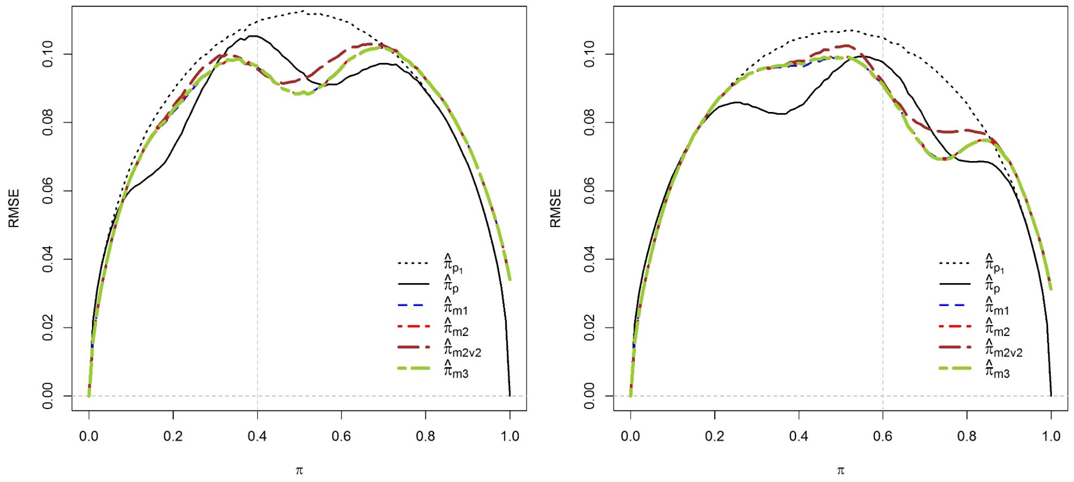
biased (Figure 3.4). The exception is the the first stage sample proportion ($\hat{\pi}_{p1}$), which is unbiased as expected. For π close to π_1, the proposed estimators ($\hat{\pi}_{m1}$, $\hat{\pi}_{m2}$, $\hat{\pi}_{m2v2}$ and $\hat{\pi}_{m3}$) are almost unbiased, while the fixed sample MLE ($\hat{\pi}_p$) shows positive bias. As π approaches 1, our proposed estimators become more negatively biased, while the bias of $\hat{\pi}_p$ approaches 0. With respect to RMSE (Figure 3.5), the proposed estimators also outperform $\hat{\pi}_p$ for values of π around π_1. They have also lower RMSE as compared to $\hat{\pi}_{p1}$.

Under H_1 the simulation mean and median are similar, and they are relatively lower in the proposed estimator as compared to $\hat{\pi}_p$ (Table 3.5). For all the proposed methods, the one-sided confidence intervals have coverage probabilities that are not less than the nominal level (95%). Their mean lower bound is similar across the proposed methods. Similar results were observed in the simulations done assuming $\pi = \pi_1 + 0.1$ (Table 3.6).



(a) Design 1 (b) Design 2

Figure 3.4: Mean bias of estimators. Design 1 is defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and 2 by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. For each value of π 50000 trials were simulated. The vertical line represents $\pi = \pi_1$.



(a) Design 1 (b) Design 2

Figure 3.5: RMSE of estimators. Design 1 is defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and 2 by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. For each value of π 50000 trials were simulated. The vertical line represents $\pi = \pi_1$.

3.3.7 Conclusion

The proposed sample space orderings are consistent with the decision rules of the designs, and so should be the corresponding overall p-values. However, some p-values are calculated using approximations and may, therefore, show conservatism.

For some values of true response probability our methods don't show improvement over the naïve MLE, nevertheless they consistently outperform the naïve MLE when the true response probability is in the neighbourhood of values that are equal to or greater than the response probability under the alternative hypothesis. It is in this region where the

TABLE 3.5: Performance measures of estimator under H_1 (i.e., $\pi = \pi_1$). The measures are the mean, median, mean bias, RMSE, first and third quartiles, and the coverage probability and mean of lower bound of the one-sided $(1-\alpha)100\%$ confidence interval. The first group of rows are for the design 1, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and the other for 2 $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. A total of 50000 trials were simulated for each design.

	True response rate: $\pi = \pi_1$				
	$\hat{\pi}_p$	$\hat{\pi}_{m1}$	$\hat{\pi}_{m2}$	$\hat{\pi}_{m2v2}$	$\hat{\pi}_{m3}$
Mean	0.4126	0.3974	0.3966	0.4008	0.3968
Median	0.4068	0.3978	0.3897	0.3975	0.3887
Mean bias	0.0126	-0.0026	-0.0034	0.0008	-0.0032
RMSE	0.1052	0.0962	0.0965	0.0956	0.0964
1^{st} quartile	0.3559	0.3385	0.3414	0.3486	0.3408
3^{rd} quartile	0.5000	0.4648	0.4677	0.4682	0.4677
Coverage probability		0.9785	0.9785	0.9785	0.9785
Lower bound		0.2685	0.2678	0.2660	0.2681
Mean	0.6057	0.5941	0.5933	0.5991	0.5935
Median	0.6027	0.5983	0.5915	0.5976	0.5904
Mean bias	0.0057	-0.0059	-0.0067	-0.0009	-0.0065
RMSE	0.0976	0.0911	0.0912	0.0920	0.0911
1^{st} quartile	0.5556	0.5502	0.5471	0.5524	0.5483
3^{rd} quartile	0.6575	0.6422	0.6383	0.6599	0.6400
Coverage probability		0.9742	0.9742	0.9742	0.9742
Lower bound		0.4723	0.4716	0.4677	0.4718

estimation becomes particularly important since the null hypothesis would likely have been rejected and the treatment effect estimate needed to plan later phase III trials. In this region the naïve MLE shows high positive bias and higher RMSE while our methods are either unbiased or negatively biased with smaller RMSE.

In general, as opposed to the naïve MLE, our proposed methods do not overestimate the response probability, they are either unbiased or negatively biased. Overestimation of treatment effect in phase II trials has been acknowledged in the literature as one of the reasons for high failure rate of drugs in phase III (see Gan et al., 2012; Kirby et al., 2012; Wang et al., 2006)

The conditional invariance principle, on which the Method 3 is based, guarantees that the type I error rate is controlled irrespective of the adaptations. Thus this method can easily be extended to other designs, including those in which adaptation rules are not pre-specified.

Table 3.6: Performance measures of estimator under $\pi = \pi_1 + 0.1$. The measures are the mean, median, mean bias, RMSE, first and third quartiles, and the coverage probability and mean of lower bound of the one-sided $(1-\alpha)100\%$ confidence interval. The first group of rows are for the design 1, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and the other for 2 $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. A total of 50000 trials were simulated for each design.

	True response rate: $\pi = \pi_1 + 0.1$				
	$\hat{\pi}_p$	$\hat{\pi}_{m1}$	$\hat{\pi}_{m2}$	$\hat{\pi}_{m2v2}$	$\hat{\pi}_{m3}$
Mean	0.5265	0.4971	0.4970	0.4931	0.4971
Median	0.5085	0.4754	0.4754	0.4754	0.4754
Mean bias	0.0265	-0.0029	-0.0030	-0.0069	-0.0029
RMSE	0.0950	0.0885	0.0884	0.0926	0.0883
1^{st} quartile	0.4746	0.4434	0.4410	0.4283	0.4425
3^{rd} quartile	0.6000	0.5737	0.5737	0.5737	0.5737
Coverage probability		0.9785	0.9785	0.9785	0.9785
Lower bound		0.3369	0.3353	0.3319	0.3369
Mean	0.7248	0.6978	0.6980	0.6925	0.6980
Median	0.7273	0.6980	0.6994	0.6994	0.6982
Mean bias	0.0248	-0.0022	-0.0020	-0.0075	-0.0020
RMSE	0.0808	0.0726	0.0722	0.0786	0.0721
1^{st} quartile	0.6761	0.6522	0.6502	0.6436	0.6509
3^{rd} quartile	0.7727	0.7462	0.7462	0.7462	0.7462
Coverage probability		0.9792	0.9792	0.9792	0.9792
Lower bound		0.5513	0.5488	0.5416	0.5515

Chapter 4

Using phase II estimates to plan phase III trials

In this chapter we discuss issues rising from the use of phase II efficacy estimates to plan phase III trials, and approaches to tackle these issues proposed it the literature. We also study, through simulations, the consequences of using estimates from adaptive phase II designs to plan phase III sample size, and we propose new approaches for adjusting these estimates. We have written and submitted for publication a paper based on this chapter.

4.1 Introduction

As discussed in Section 2.1, the clinical drug development is mainly done in three phases, phase I, phase II and phase III. The knowledge gained in clinical trials of a particular phase is often used to plan trials of subsequent phases. That is the case with successful phase II clinical trials in which, among others aspects, the effect size estimates are used to plan the sample size of the related phase III trials. Due to small sample sizes, selections bias and other factors, phase II estimates are often biased and imprecise, resulting in inadequately powered phase III trials.

A high failure rate of phase III trials has been reported in the literature. In general, approximately 40% of phase III trials fail (De Martini, 2013). In oncology the situation is even worse, for instance, Gan et al. (2012) found that of 253 phase III randomized clinical trials (RCTs) evaluating systemic therapy in adult cancer patients published in 10 journals from January 2005 to December 2009, 62% did not achieve statistically significant results. They noted that the actual magnitude of benefit achieved in a clinical trial is nearly always less than what was predicted at the time the trial was designed, and that investigators consistently make overly optimistic assumptions regarding treatment benefits when designing RCTs. One of the reasons for this optimism might be the over-estimation of the treatment effect in phase II trials, acknowledged by Wang et al. (2006) and Kirby et al. (2012), who proposed methods for adjusting (discounting) the phase II treatment effect estimate when employing it to plan the sample size of a phase III trial.

We have written a paper based on this chapter. The paper was still under submission to a scientific journal by the time of the writing of this dissertation. The submitted version is in Appendix E.

4.2 Dealing with bias and imprecision

The problem of biased and imprecise efficacy estimates from phase II trials has been acknowledged in the literature, and some authors (e.g., Burke et al., 2014; Chuang-Stein and Kirby, 2017; De Martini, 2011a,b, 2013; Kirby et al., 2012; Wang et al., 2006) have discussed ways to account for it when planning the subsequent phase trials. The common approach of directly using these efficacy estimates to calculate sample size often results, as noted above, in inadequately powered studies. To take into account the bias and variability of phase II data for planing phase III trials, Wang et al. (2006), Kirby et al. (2012) and De Martini (2013) discussed different *conservative sample size estimation* (CSSE) strategies, which aim at controlling the success probability (SP; i.e., power). To define these strategies, concepts of launch criteria for phase III, variability of sample size estimates, and averaged SP of phase III (average power) are used. Launch criteria define requirements under which a phase III study is initiated. The most common is the *statistical significance* criterion, under which phase III is launched if phase II results are statistically significant. Other criteria include the *clinical relevance*, under which phase III is launched if the phase II effect estimate is larger than a specified effect size deemed to be of a clinical relevance, and the *maximum sample size* criterion, where phase III is launched if the estimated sample size does not exceed a certain threshold. This sample size threshold might be defined on the basis of budget constraints or patient availability. The CSSE strategies can be grouped into frequentist and Bayesian frameworks. In general, the frequentist approach to CSSE consists in using a conservative value ($\hat{\theta}_f$) of the phase II effect size estimate ($\hat{\theta}$) used to determine the phase III sample size. This can be achieved by either subtracting a certain amount from $\hat{\theta}$, e.g., one standard error (i.e., $\hat{\theta}_f = \hat{\theta} - \mathrm{SE}(\hat{\theta})$), or by applying a discounting factor $f \in]0, 1]$ (i.e., $\hat{\theta}_f = \hat{\theta} \times f$). The Bayesian CSSE considers the posterior distribution of the effect θ instead of its point estimate $\hat{\theta}$. This strategy puts a probability mass around the observed phase II effect and computes the averaged SP at a given sample size, giving the Bayesian estimate of the SP. Then the phase III sample size estimate is the minimum sample size whose Bayesian SP exceeds a certain desired power.

Where many similar phase II trials on the same therapy exist, meta-analytic approaches can also be used to better plan subsequent phase III trials. For instance, in the context of randomized phase II trials with binary endpoints, Burke et al. (2014) deemed the meta-analysis using a Bayesian random effects logistic regression model to be the most appropriate. With the model, predictions that inform phase III decision can be made, namely the probability that the therapy will be truly effective in a new trial, and the probability that, in a new trial with a given sample size, the 95% credible interval for the odds ratio will be entirely in favour of the therapy. They also argue in favour of using sceptical prior distributions to reduce optimism of phase II trials in order to make more realistic predictions.

4.3 Using estimates from oncology phase II adaptive trials

Despite the multitude of approaches proposed in the literature, as discussed above, the question of how to appropriately employ phase II effect estimates to plan phase III trials still remains a challenge. The frequentist CSSE strategies that encourage discounting the effect estimate beforehand do not offer "one size fits all" guidelines on the amount of effect that ought to be discounted. The amount of effect to be discounted is more likely to depend on the circumstances and characteristics of a specific trials. In fact, authors that provide some guidelines they do so based on empirical studies under restrictive assumptions and, therefore, these guidelines may not be applicable in other scenarios. The Bayesian CSSE strategies also suffer from similar problems. The prior distribution is the "Achilles' heel" in these strategies. The choice of an adequate conservative prior distribution can be a daunting task for which universally valid guidelines are hard to establish.

These difficulties are even more pronounced in oncology phase II adaptive trials where there are no well established estimation methods in first place. Having estimators that account for adaptiveness of these designs and that suffer less from upwards bias can alleviate the problem. Some of such estimators are the ones we proposed in Section 3.3.

4.3.1 Simulation study

Here we evaluate through simulation studies the consequences, in terms of power, of using the effect estimate from oncology phase II adaptive design trials to plan sample size of a related phase III trial. We consider the recently proposed oncology phase II two-stage single-arm adaptive designs with binary endpoint, in which the second stage sample size is a pre-defined function of the first stage's number of responses (successes). An example of such designs is given in Table 3.3 of Chapter 3, and more examples and details can be found in Englert and Kieser (2013) and Shan et al. (2016a,b). Different estimators are used. The naïve (fixed-sample) maximum likelihood estimator, which is more likely to be employed for adaptive design for which no specific estimator has been proposed, and our proposed estimates (see Chapter 3, Section 3.3). For simplicity, we consider two-arm phase III RCTs also with binary endpoint. Although a survival endpoint is commonly used in oncology phase III trials, there are some types of cancer for which the response rate is a suitable endpoint. The objective response rate (ORR), as defined by RECIST guidelines (Eisenhauer et al., 2009), is the most commonly used binary endpoint in oncology trials. ORR has been used as the primary endpoint in 40% of advanced breast cancer phase III trials published between January 1998 and December 2007 (Saad et al., 2010).

4.3.1.1 Trial designs and estimation methods

The phase II designs and estimation methods we are considering here are given in more details in Section 3.3. Here we give only a short summary.

We consider oncology phase II adaptive designs similar to those proposed by Englert and Kieser (2013) and Shan et al. (2016a,b). These are binary endpoint single-arm two-stage designs, testing at type I error rate α and type II error rate β the null (H_0) versus the alternative (H_1) hypothesis, $H_0 : \pi \leq \pi_0$ vs $H_1 : \pi \geq \pi_1$, where π_0 is the maximum response rate considered to be uninteresting and π_1 is the minimum desirable response rate, with $\pi_1 > \pi_0$. In the first stage, these designs are characterized by the sample size, n_1, and the futility and efficacy boundaries, l_1 and u_1 ($u_1 > l_1$), which are fixed, and in the second stage by the sample size, $n_2(x_1)$, the conditional error function, $D(x_1)$, and the corresponding decision boundary, $l(x_1)$, which are pre-specified functions of the number of responses, x_1, observed in the first stage. At the interim analysis, the trial is stopped with no rejection of H_0 if $x_1 \leq l_1$ or with rejection of H_0 if $x_1 \geq u_1$. Otherwise the trial proceeds to the second (final) stage, at which H_0 is rejected if $p_2 \leq D(x_1)$ or, equivalently, $x > l(x_1)$, where p_2 is the second stage p-value and x is the total number of responses (i.e., $x = x_1 + x_2$).

We assume that a successful phase II trial will be followed by a single-stage randomized parallel-group phase III clinical trial with binary endpoint, similar to the design described by Halabi (2008). The trial tests the null hypothesis that the proportion of response (success) in the control and treatment groups, π_c and π_t, are equal, i.e., $H_0 : \pi_c = \pi_t$, versus $H_1 : \pi_c \neq \pi_t$.

The effect estimate from the phase II trial that is to be used for calculating the required sample size of the subsequent phase III trial will be obtained using the following estimation methods. The naïve MLE is calculated as

$$\hat{\pi}_{nml} = x/n.$$

As it can be seen, this estimate ignores the adaptive nature of the design, treating the data as if they were from a single-stage non-adaptive trial. Let (m, x_1, x) be the outcome of atrial that stopped at stage m with first stage's and total number of responses, x_1 and x. Our estimate (see more details in Section 3.3), that takes into account the adaptiveness of the design, is defined as

$$\hat{\pi}_m = \{\tilde{\pi}_0 : Q(\tilde{\pi}_0) = \Pr_{\tilde{\pi}_0}\left((M, X_1, X) \succeq (m, x_1, x)\right) = 0.5\},$$

where Q is the overall p-value based on sample space orderings we proposed, and which can be calculated as

$$Q = \begin{cases} 1 - B(x_1 - 1, n_1, \tilde{\pi}_0) & \text{if } m = 1 \\ 1 - B(u_1 - 1, n_1, \tilde{\pi}_0) + \sum\limits_{X_1 = l_1 + 1}^{u_1 - 1} b(X_1, n_1, \tilde{\pi}_0) \Pr_{\tilde{\pi}_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2)\right] & \text{if } m = 2 \end{cases}$$

where $B(x, n, \pi)$ and $b(x, n, \pi)$ are the binomial cumulative distribution function and probability mass function with x successes, n trials and success probability π. As discussed in Section 3.3, we modify the classical stage-wise sample space ordering to take into account the design's adaptation rule by defining the function $\delta(x_1, x_2)$ that somehow incorporates the rejection boundary of the trial outcome. We use three different methods to define $\delta(x_1, x_2)$. In the first method δ is defined using x as $\delta(x_1, x_2) = x - l(x_1)$, in the second method defined using the second stage p-value as $\delta(x_1, x_2) = D(x_1) - p_2(x_2)$ and, finally, in the third method defined as $\delta(x_1, x_2) = 1 - C(p_{1b}, p_2)$, where C is the weighted inverse normal combination function represented as $C(p_1, p_2) = 1 - \Phi\left[w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)\right]$, with

$$w_1 = \sqrt{\frac{n_1}{n_1 + n_2(x_1)}},\ w_2 = \sqrt{\frac{n_2(x_1)}{n_1 + n_2(x_1)}}$$

and

$$p_{1b}(x_1) = 1 - \Phi\left\{\frac{\Phi^{-1}(1 - c) - w_2 \Phi^{-1}\left[1 - D(x_1)\right]}{w_1}\right\}.$$

Following the same notation used in Section 3.3, we denote the estimated response probability by $\hat{\pi}_{m1}$ for δ defined in terms of number of responses and rejection boundary, $\hat{\pi}_{m2}$ and $\hat{\pi}_{m2v2}$ for for δ defined in terms of second stage p-value and conditional error function (here two different approaches are used to calculate $\Pr_{\tilde{\pi}_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2)\right]$, hence the two notations), and $\hat{\pi}_{m3}$ for δ defined by the inverse normal combination function.

4.3.1.2 Simulation set-up

Acknowledging the over-estimation of phase II treatment effect and its negative consequences on planing phase III trials discussed in the literature, we conducted a simulation study to assess the extent of this problem in oncology trials. The simulation set-up was as follows. Pick a particular phase II adaptive design testing $H_0 : \pi \leq \pi_0$ vs $H_1 : \pi \geq \pi_1$ at a specific type I and type II error rates α and β as described above. Simulate K trials assuming a specific true response probability π. For the simulated trials in which the null hypothesis was rejected, get the estimate of π, denoted $\hat{\pi}$. Assume that the subsequent phase III trial has a similar endpoint and, as described in the previous section, it tests $H_0 : \pi_c = \pi_t$ versus $H_1 : \pi_c \neq \pi_t$. Assume further that the desired type I and type II error rates α' and β', respectively, and that $\pi_c = \pi_0$. Calculate the required sample size, N, to detect the effect size of magnitude $\hat{\pi} - \pi_c$ with power of $1 - \beta'$. Using N, calculate what would be the attained power to detect the true effect size, $\pi - \pi_c$ (recall that π is the response probability under which the trials were simulated, and $\hat{\pi}$ is its estimate). The different estimators mentioned were used to obtain $\hat{\pi}$. Different values of retention factor f was applied before computing N. The factor f was proposed by Kirby et al. (2012) in their multiplicative adjustment approach, under which $f \in [0, 1]$ is applied to the estimate of phase II treatment effect $\hat{\pi}$ to obtain a multiplicatively adjusted treatment effect estimate $\hat{\pi}_f = \hat{\pi} \times f$. The adjusted effect estimate $\hat{\pi}_f$ can be viewed as the result of discounting $\hat{\pi}$ by $100(1 - f)\%$ and it is meant to be used for planing phase III trials. Unlike Kirby et al. (2012), we do not define a launch criterion based on a threshold of $\hat{\pi}_f$, instead we assume that a phase III trial is launched whenever the null hypothesis is rejected in the phase II trial. Therefore, we discard the simulated phase II trials in which H_0 is not rejected. This means that the phase III power we are aiming at is conditional on rejection of H_0 in phase II. This might be the most important scenario since in practice only successful phase II trials are likely to be used in planning subsequent phase III trials.

For power and sample size calculation of the phase III trial described above, we use the two-sample test for proportion described by Ahn et al. (2014). The power is approximated by

$$\Phi\left(\frac{\pi_t - \pi_c}{\sqrt{\pi_t(1-\pi_t)/N_t + \pi_c(1-\pi_c)/N_c}} - z_{1-\alpha/2}\right),$$

and the sample size $N = N_c + N_t$ needed to achieve a power of $1-\beta$ obtained by solving the equation

$$\frac{\pi_t - \pi_c}{\sqrt{\pi_t(1-\pi_t)/N_t + \pi_c(1-\pi_c)/N_c}} - z_{1-\alpha/2} = z_{1-\beta},$$

where Φ and z_u are the standard normal cumulative distribution function and u-quantile, and N_t and N_c are the sample sizes for the treatment and control groups, respectively. In the simulations, we assume equal size groups, hence

$$N_t = N_c = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\pi_t - \pi_c\right)^2}\left[\pi_t(1-\pi_t) + \pi_c(1-\pi_c)\right].$$

The simulation scenarios are as follow. Two phase II adaptive designs of Englert and Kieser (2013) were used. One design, let's call it design 1, is defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and the other, design 2, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. The true response rate π varied from $\pi_0 + 0.1$ to $\pi_1 + 0.3$ by increments of 0.01. When estimating the phase III sample size, we assumed a type I error rate of 5% and a power of 90%, and the retention factor f varied from 0.5 to 1 by increments of 0.01. Note that $f = 1$ means no effect is retained, i.e., the original estimate is used. For each scenario 50000 simulation runs were done.

All the simulations and computation were done using the statistical programming language R (R Core Team, 2017).

4.3.1.3 Results

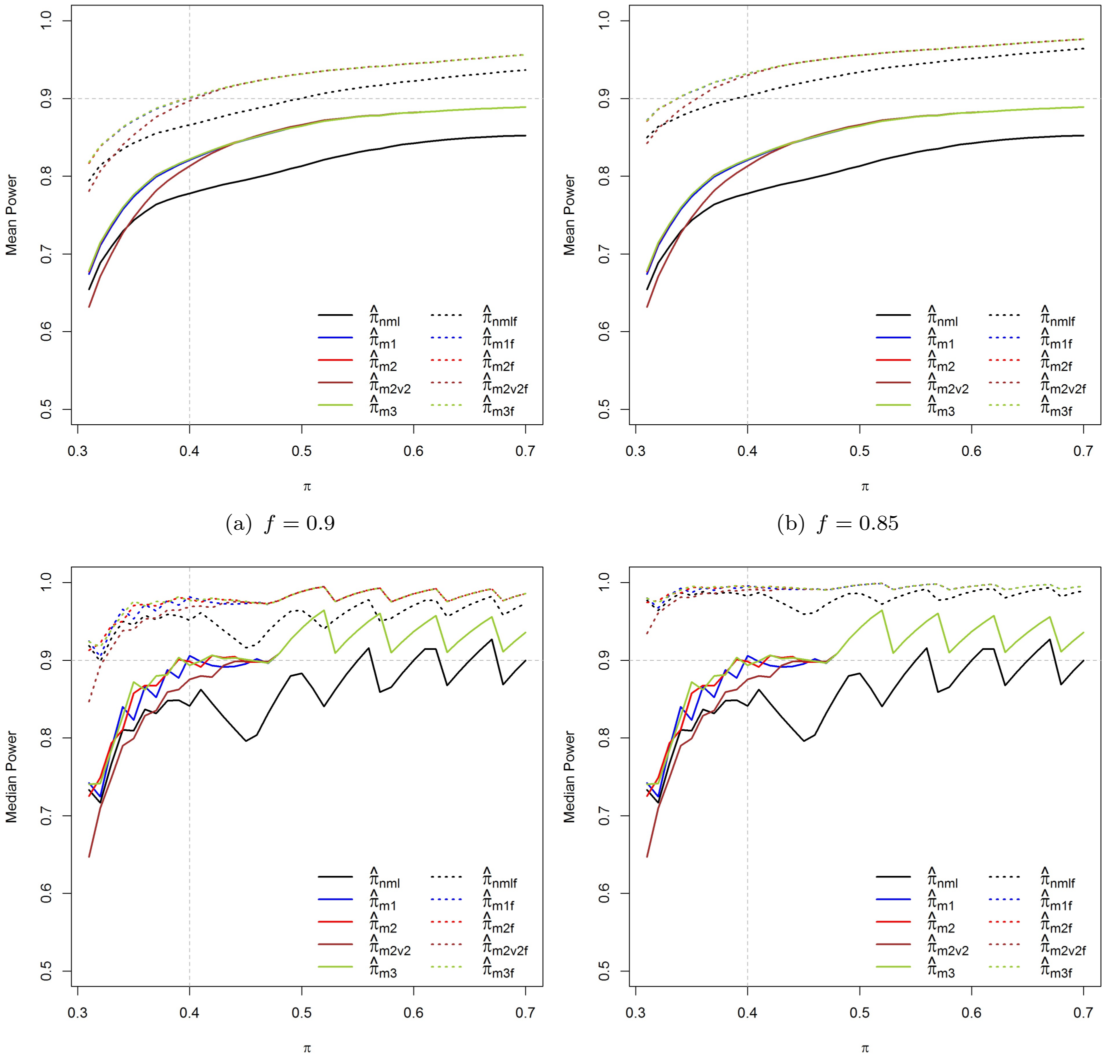
Simulations results for $\pi = \pi_1$ and $\pi = \pi_1 + 0.1$, and $f = 1$, $f = 0.9$ and $f = 0.85$ are shown in Tables 4.1 and 4.2 for designs 1 and 2, respectively. The results for all the values of π ranging from $\pi_0 + 0.1$ to $\pi_1 + 0.3$ by increments of 0.01 are plotted in Figures 4.1 and 4.2. The results show the power attained in phase III trial whose sample size was planned using effect estimate from phase II adaptive design. The effect estimates were calculated using estimation methods described in Section 4.3.1.1, namely the estimates that take into account the adaptiveness of the design, $\hat{\pi}_{m1}$, $\hat{\pi}_{m2}$, $\hat{\pi}_{m2v2}$ and $\hat{\pi}_{m3}$, and the naïve estimate, $\hat{\pi}_{nml}$ (fixed-sample maximum likelihood estimate). These results show that, in general, when using phase II effect size estimates to plan the phase III sample size, estimates from the adaptive estimation methods (i.e., $\hat{\pi}_{m1}$, $\hat{\pi}_{m2}$, $\hat{\pi}_{m2v2}$ and $\hat{\pi}_{m3}$) yield better power as compared to the naïve estimate ($\hat{\pi}_{nml}$). For instance, when the true response rate is equal to the value under H_1, $\hat{\pi}_{m1}$ in design 1 (Table 4.1) yielded a mean power of 82.1% and median power of 90.6%. The median power is very close to the target value of 90%, and the average power is below by about 8%. The mean and median power for the naïve MLE $\hat{\pi}_{nml}$ are even smaller, namely 77.8% and 84.1%. For the naïve MLE, the retention factor of 0.9 (i.e., discounting 10% of the effect estimate) seems to be suitable to bring the power closer to the target value, resulting in the mean power of 86.7% and median power of 95.2%. This retention factor also seems to guarantee that the mean and the median power of the adaptive estimates are both not less than the target value. Regarding the behaviour for different values of the true response rate (see Figures 4.1 and 4.2), all the estimates yield under-powered phase III trials when π is less than π_1. But as π increases, the adaptive estimates yield higher power than the naïve one. The power among the adaptive estimates also becomes more homogeneous as π gets bigger than π_1.

Table 4.1: Mean and median power of phase III trials planned using phase II design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$. The target power of the phase III trial is of 90%.

		$f = 1$		$f = 0.9$		$f = 0.85$	
π	$\hat{\pi}$	Mean	Median	Mean	Median	Mean	Median
π_1	$\hat{\pi}_{nml}$	0.7779	0.8414	0.8660	0.9515	0.9035	0.9828
	$\hat{\pi}_{m1}$	0.8207	0.9060	0.9009	0.9818	0.9321	0.9958
	$\hat{\pi}_{m2}$	0.8221	0.8986	0.9013	0.9789	0.9322	0.9949
	$\hat{\pi}_{m2v2}$	0.8131	0.8755	0.8969	0.9690	0.9300	0.9910
	$\hat{\pi}_{m3}$	0.8219	0.8936	0.9013	0.9769	0.9322	0.9941
$\pi_1 + 0.1$	$\hat{\pi}_{nml}$	0.8131	0.8834	0.9004	0.9646	0.9340	0.9863
	$\hat{\pi}_{m1}$	0.8648	0.9416	0.9315	0.9887	0.9556	0.9970
	$\hat{\pi}_{m2}$	0.8648	0.9416	0.9316	0.9887	0.9556	0.9970
	$\hat{\pi}_{m2v2}$	0.8660	0.9416	0.9317	0.9887	0.9556	0.9970
	$\hat{\pi}_{m3}$	0.8648	0.9416	0.9316	0.9887	0.9556	0.9970

Table 4.2: Mean and median power of phase III trials planned using phase II design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 20)$. The target power of the phase III trial is of 90%.

		$f = 1$		$f = 0.9$		$f = 0.85$	
π	$\hat{\pi}$	Mean	Median	Mean	Median	Mean	Median
π_1	$\hat{\pi}_{nml}$	0.7884	0.8623	0.9105	0.9912	0.9495	0.9998
	$\hat{\pi}_{m1}$	0.8232	0.9050	0.9354	0.9971	0.9670	1.0000
	$\hat{\pi}_{m2}$	0.8250	0.9003	0.9356	0.9967	0.9670	1.0000
	$\hat{\pi}_{m2v2}$	0.8063	0.8752	0.9283	0.9934	0.9649	0.9999
	$\hat{\pi}_{m3}$	0.8248	0.8982	0.9356	0.9965	0.9670	1.0000
$\pi_1 + 0.1$	$\hat{\pi}_{nml}$	0.8075	0.8346	0.9341	0.9731	0.9699	0.9957
	$\hat{\pi}_{m1}$	0.8655	0.9005	0.9599	0.9913	0.9827	0.9993
	$\hat{\pi}_{m2}$	0.8656	0.9013	0.9599	0.9915	0.9827	0.9993
	$\hat{\pi}_{m2v2}$	0.8676	0.9013	0.9599	0.9915	0.9827	0.9993
	$\hat{\pi}_{m3}$	0.8656	0.9002	0.9599	0.9913	0.9827	0.9993



(a) $f = 0.9$

(b) $f = 0.85$

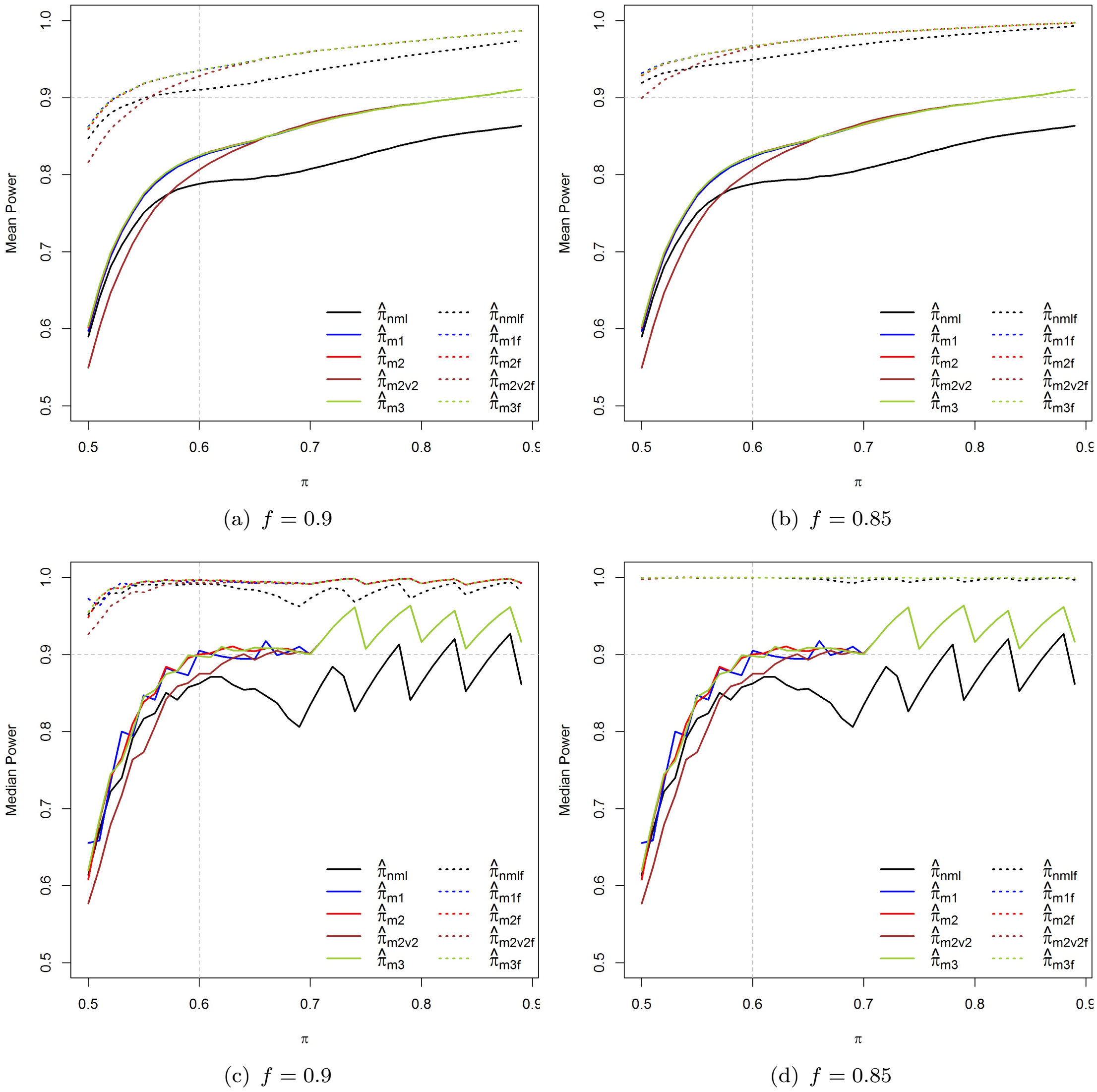
(c) $f = 0.9$

(d) $f = 0.85$

FIGURE 4.1: Mean and median power of phase III trials planned using phase II design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$. The solid lines are for the unadjusted effect estimates, and the dashed lines for the adjusted ones (i.e., estimates multiplied by the retention factor f). For each value of π 50000 trials were simulated. The vertical line represents $\pi = \pi_1$ and the horizontal one represents the target power (90%).

4.3.1.4 Conclusion

We have studied via simulations the impact on power when using effect estimates from phase II adaptive oncology trials to plan phase III trials. Results showed that using the estimators that accounts for the adaptiveness of the design yield better results than the naïve estimates, as the power of the resulting phase III trials is higher than that of trials planned using naïve estimates. However, as far as the mean power is concerned, none of the estimators yields the target (nominal) phase III power. That means that effect retention is necessary irrespective of which estimator is employed, even when the



(a) $f = 0.9$

(b) $f = 0.85$

(c) $f = 0.9$

(d) $f = 0.85$

FIGURE 4.2: Mean and median power of phase III trials planned using phase II design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 20)$. The solid lines are for the unadjusted effect estimates, and the dashed lines for the adjusted ones (i.e., estimates multiplied by the retention factor f). For each value of π 50000 trials were simulated. The vertical line represents $\pi = \pi_1$ and the horizontal one represents the target power (90%).

estimator is unbiased. This is due the fact that the power of the phase III trial is a highly non-linear function of the phase II effect estimate and we consider the power conditional to rejection in phase II. The extent of retention will vary from one estimator to another, with better performing estimators requiring less reduction of the effect estimate. On the other hand, if the median power is taken as the metric, the median unbiased estimator accounting for adaptiveness yields the target power and, then the effect adjustment approach proposed by Wang et al. (2006) and Kirby et al. (2012) becomes unnecessary. We noticed that the empirical distribution of the power is skewed, and this might be the reason for these differences between the mean and the median. It also raises the question of which of the two is the best measure of central tendency.

The simulation results in section 3.3.6.1 showed that for values of the true response rate lower than the hypothesized one (i.e., $\pi < \pi_1$) all the estimators (the naïve and their proposed ones) underestimate π. For higher true response rates, the naïve estimator overestimates π while the others are either unbiased or underestimate. Therefore we would expect that in our simulations all the estimators result in overpowered phase III trials for $\pi < \pi_1$, which is not the case. This counter-intuitiveness is due to the fact that here we only consider for the power calculations estimates from trials that led to the rejection of the null hypothesis, and these are more likely to be few outlying trials. When $\pi < \pi_1$ the majority of trials fail to reject H_0 and, hence are excluded. This results in a overestimated π, which in turn leads to underpowered phase III trials.

In this simulation study, we assumed that the true treatment effect in phase III is the same as in phase II, and that the phase III sample size is estimated based on the phase II treatment effect estimate. We think this is likely to be the case if the phase II trial is the only source of information (regarding treatment effect) for planning a subsequent phase III trial. Then the focus of the investigators will be on how to adjust this single phase II effect estimate in order to properly power the phase III trial. However, there will be cases where multiple sources of information are available to aid the planning of phase III. In such situations, other approaches would be more appropriate than the one we took here. One example is the case where data from multiple similar phase II trials are available. In this case, meta-analytic approaches (e.g., Burke et al., 2014) are more useful.

4.4 New approaches for adjusting phase II estimates

The results from the simulation study above suggest that it is almost always necessary to adjust phase II estimates before using them to plan phase III trials. However, as mentioned before, despite the existence of many approaches to do so (see Section 4.2), in practice the adjustment of these estimates still remains a challenge. The frequentist conservative sample size estimation strategies lack clear guidelines on the amount of effect to discount. And the Bayesian strategies rely on the choice of conservative prior distribution, for which universally valid guidelines are difficult to establish.

Here we present new approaches on how to estimate the Kirby et al. (2012)'s multiplicative adjustment factor to be applied to Phase II treatment effect estimates before employing them to plan the sample size of the related Phase III clinical trials, based on the observed data. Alternatively, the approaches can also be used to estimate the adjustment factor to be applied to the phase III sample size planned using unadjusted phase II efficacy estimates. These approaches are based on parametric bootstrapping.

4.4.1 Method 1

Suppose that a phase II trial is conducted with the intent of testing the hypothesis $H_0 : \theta \leq \theta_0$ versus $H_0 : \theta \geq \theta_1$, where θ is the efficacy parameter, and θ_0 and θ_1 its values considered to be of no clinical interest and of clinical interest, respectively. Let Y be the observed data, which is drawn from a parametric distribution $\mathcal{F}(\theta)$. Assume that Y led to the rejection of H_0 and, therefore, it is decided that the treatment under study is worthy of further investigations in phase III. The estimate of θ, $\hat{\theta}$, is obtained from Y and, assuming specific type II error rate (α) and power ($1-\beta$) constraints, the minimum required sample size, $\hat{n} = \hat{n}(\hat{\theta}, \alpha, \beta)$, for the related subsequent phase III trial is calculated. Then phase II trials τ_i, $i = 1, \ldots, m$, are simulated, with data points drawn from the distribution $\mathcal{F}(\hat{\theta})$. The simulated trials are similar in design to the actual (conducted) phase II trial. Let J be the index set of all the simulated trials in which the null hypothesis was rejected, and $|J|$ its cardinality ($|J| \leq m$) . For each simulated trial τ_j, $j \in J$, the estimate of $\hat{\theta}$, denoted by θ^*_j, is obtained and, assuming the same type I error and power constraints, the minimum required phase III sample size, $n^*_j = n^*_j(\theta^*_j, \alpha, \beta)$, is calculated. Then the individual estimates of the multiplicative adjustment factors for the effect size and sample size, f_j and ρ_j, are calculated as $f_j = \hat{\theta}/\theta^*_j$ and $\rho_j = \hat{n}/n^*_j$. The average values of these factors are taken to be their final estimates, i.e., $f = \frac{1}{|J|}\sum_{j \in J} f_j$ and $\rho = \frac{1}{|J|}\sum_{j \in J} \rho_j$. Therefore, the adjusted efficacy to be used in planing phase III trials is $\theta_f = \hat{\theta} \times f$. Alternatively, the adjusted sample size $n_\rho = \hat{n} \times \rho$ could be used. Since the sample size is a non-linear function of the effect estimate, the two adjustment approaches will, in general, lead to different final (adjusted) sample sizes and, consequently, to differences in power. We will see this in the simulation study presented below.

4.4.2 Method 2

Here we present an alternative, more direct method to estimate the sample size adjustment factor ρ. We use the same notation as in the previous method (Section 4.4.1). Let $\text{pwr}^*_j = \text{pwr}(n^*_j, \hat{\theta}, \alpha)$ be the power a phase III trial would attain, with sample size n^*_j if the true and the observed efficacy estimate (θ and $\hat{\theta}$) would coincide (see previous sections for details). The sample size multiplicative adjustment factor is calculated such that the expected value of pwr^* is equal to the target (desired) power assuming (as an approximation) that $\hat{\theta}$ and θ coincide, i.e.,

$$\rho = \left\{ \tilde{\rho} | E_{\hat{\theta}} \left[\text{pwr}(\tilde{\rho} n^*_j, \hat{\theta}, \alpha) \right] = 1 - \beta \right\}.$$

The adequate value of ρ can be found using numerical root finding.

4.4.3 Simulation study

We study, using simulations, the behaviour of our new approaches. We are interested in knowing to which extent the adjustment using our proposed methods results in properly powered phase III trials. We also study, in addition, the variability of the resulting adjustment factors.

The simulations are as follows. Assuming a specific true treatment effect, we simulate phase II trials. For each simulated trial in which the null hypothesis was rejected, we calculate the effect estimate and we apply our approaches (see Section 4.4) to obtain the adjustment factors which we then use to estimate the phase III sample size. Then, given the true treatment effect (under which trials were simulated), we calculate the power that would be attained with the adjusted sample size estimate. The phase II and phase III designs, the hypotheses, the type I and type II error rates, and the sample size and power calculation are the same as those used in the simulation study of Section 4.3.1. The phase II design has a binary endpoint, therefore, the distribution function $\mathcal{F}$ described in Section 4.4 is binomial with parameter π (response probability). The phase II designs are simulated under the alternative hypothesis, i.e., $\pi = \pi_1$. We use the two designs of Section 4.3.1.2, i.e., the design 1 defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$ and the design 2 defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. We assumed a type I error rate of 5% and a power of 90% when estimating the phase III sample size. The phase II effect estimate is obtained using also the same estimators as in Section 4.3.1.

For each scenario, 5000 trials were simulated, and for each simulated trial in which the null hypothesis was rejected, 5000 bootstrap samples (trials) were obtained (for calculating the adjustment factors).

4.4.3.1 Results

Table 4.3 shows the results of the simulations. As it can be seen, the naïve MLE requires that a higher amount of effect is discounted as compared to the others estimators that take into account the adaptiveness of the designs. This is seen by the fact that, for this estimator (MLE), the effect adjustment factor (f) is lower and, equivalently, the sample size adjustment factor (ρ) is higher. Method 2 yields higher ρ as compared to Method 1. Both methods improve the power of the phase III trial. For instance, for the naïve MLE the average power without adjustment is 78.8%, which is increased to 80.7% and 85.2% by the Method 1 using f and ρ respectively, and to 86.6% using the Method 2. Method 2 is the best performer as it yields average power that is closer to the target (90%). The median power is higher than the average in all cases (including when no adjustment is applied), and values that a closest to the target power are attained by using f in Method 1. We found out that empirical distributions of the adjustment factors and the power in this simulation (results not shown) are skewed, resulting in differences of the average and median values. In general the results from both methods are similar across different estimators and designs, especially the results from Method 2.

4.4.3.2 Conclusion

The results in this simulation showed that, when planning phase III sample size based of phase II effect estimates, adjustments are almost always necessary. However, the extent of adjustment will differ depending on the estimator that is employed to get phase II effect estimate. In our specific case, the estimators that account for the adaptive nature of the designs require less correction. Our proposed adjustment methods show improvements in power, and the results are consistently similar for different estimators and design scenarios. This suggests that our approach may perform well in other design scenarios and with different estimators. When the metric of interest is the mean power, Method 2 is more preferable over Method 1, especially in cases where the power distribution is known to be skewed.

TABLE 4.3: Multiplicative adjustment factors for effect size and sample size (f and ρ), and the corresponding attained power in phase III trial. The first group rows correspond to the phase II design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$ and the other group to the design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 20)$. The target for power is 90%.

	Adjustment factor						Power							
	Method 1				Method 2		No adjustment		Method 1				Method 2	
	f		ρ		ρ		pwr$[n(\hat{\theta}), \theta, \alpha]$		pwr$[n(f\hat{\theta}), \theta, \alpha]$		pwr$[\rho n(\hat{\theta}), \theta, \alpha]$		pwr$[\rho n(\hat{\theta}), \theta, \alpha]$	
Est.	Mean	SD	Mean	SD	Mean	SD	Mean	Median	Mean	Median	Mean	Median	Mean	Median
$\hat{\pi}_{nml}$	0.9515	0.0337	1.5906	0.3673	1.7311	0.3930	77.8%	84.1%	80.7%	88.7%	85.2%	95.0%	86.6%	96.6%
$\hat{\pi}_{m1}$	0.9825	0.0472	1.4062	0.3477	1.5128	0.3564	82.1%	90.6%	81.4%	91.6%	85.9%	96.3%	87.3%	97.4%
$\hat{\pi}_{m2}$	0.9834	0.0477	1.4143	0.3713	1.5260	0.3846	82.3%	89.9%	81.5%	90.7%	85.9%	95.9%	87.3%	97.2%
$\hat{\pi}_{m2v2}$	0.9799	0.0566	1.4271	0.3824	1.5477	0.4086	81.3%	87.5%	80.4%	89.0%	85.1%	94.5%	86.7%	96.0%
$\hat{\pi}_{m3}$	0.9835	0.0474	1.4128	0.3694	1.5238	0.3820	82.3%	89.4%	81.4%	90.2%	85.9%	95.6%	87.3%	97.0%
$\hat{\pi}_{nml}$	0.9658	0.0170	1.5910	0.3202	1.7063	0.3080	78.8%	86.2%	82.5%	91.2%	86.2%	96.2%	87.5%	97.6%
$\hat{\pi}_{m1}$	0.9841	0.0259	1.4049	0.3201	1.4917	0.3107	82.2%	90.5%	82.6%	92.3%	86.3%	96.4%	87.5%	97.4%
$\hat{\pi}_{m2}$	0.9845	0.0260	1.4113	0.3386	1.5018	0.3332	82.5%	90.0%	82.7%	91.7%	86.4%	96.1%	87.6%	97.3%
$\hat{\pi}_{m2v2}$	0.9818	0.0341	1.4495	0.3941	1.5723	0.4157	80.3%	86.6%	80.3%	89.2%	84.4%	94.3%	86.0%	96.0%
$\hat{\pi}_{m3}$	0.9845	0.0258	1.4094	0.3364	1.4991	0.3304	82.4%	89.8%	82.7%	91.6%	86.4%	96.0%	87.6%	97.2%

Chapter 5

Summary and conclusions

The main achievements of this dissertation are the new inference methods for oncology phase II adaptive designs and the new methods for adjusting the phase II estimates when using them to plan the sample size of phase III clinical trials.

In the new inference methods for the adaptive designs, we started by proposing new sample space orderings, from which we derived methods for calculating p-value and then methods for interval and point estimation. Simulation studies showed good performance of our methods, with p-values being in concordance with the designs' decision rules, confidence intervals attaining the nominal coverage probability, and point estimates having lower bias and mean square error as compared to the fixed-sample maximum likelihood estimator. These inference methods can serve as useful tools for researchers who may decide to run trials following adaptive designs which are improved versions of the popular Simon-like classical group-sequential designs, or even encourage the use of such designs in first place.

Our proposed adjustment methods are based on bootstrapping, and provide estimates of adjustment factors, given the observed phase II data, for the phase II effect estimate or for the phase III sample estimate in order to get an appropriately powered phase III trial. Simulation studies showed that these methods perform well and consistently under different efficacy estimators and design configurations. These methods may also help investigators to efficiently plan phase III sample size based on phase II data, since the existing approaches lack clear guidelines. They can in principle be applied to any type of endpoint and estimate (frequentist or Bayesian).

Appendix A

Oncology phase II designs: references

Table A.1: Single-arm designs

Single-arm single-stage designs:
Jung (2013); Fleming (1982); Storer (1992); Conaway and Petroni (1995); Conaway and Petroni (1996); Sargent et al. (2001); A'Hern (2004); London and Chang (2005); Jin (2007); Stallard and Cockey (2008); Zhong and Zhong (2013); Zhong (2012); Sylvester (1988); Hilden (1990); Brunier and Whitehead (1994).

Single-arm sequential designs:
Thall and Simon (1994); Lee and Liu (2008); Johnson and Cook (2009).

Single-arm classical group-sequential designs:
Gehan (1961); Schultz et al. (1973); Lee et al. (1979); Fleming (1982); Chang et al. (1987); Simon (1989); Therneau et al. (1990); Storer (1992); Ensign et al. (1994); Chen et al. (1994); Bryant and Day (1995); Conaway and Petroni (1995); Conaway and Petroni (1996); Chen (1997); Herndon (1998); Hanfelt et al. (1999); Zee et al. (1999); Lin and Chen (2000); Jung et al. (2001); Sargent et al. (2001); Panageas et al. (2002); Case and Morgan (2003); A'Hern (2004); Lu et al. (2005); London and Chang (2005); Jin (2007); Ye and Shyr (2007); Ayanlowo and Redden (2007); Chi and Chen (2008); Chen and Shan (2008); Lin et al. (2008); Stallard and Cockey (2008); Kocherginsky et al. (2009); Mander and Thompson (2010); Kunz and Kieser (2011); Tan and Xiong (2011); Ray and Rai (2011, 2012); Chen and Chi (2012); Zhong and Zhong (2013); Zhong (2012); Ray and Rai (2013); Chen and Lee (2013); Whitehead (2014); Kwak and Jung (2014); Poulopoulou et al. (2014); Song (2015); Lai and Zee (2015); Herson (1979); Heitjan (1997); Tan and Machin (2002, 2006); Sambucini (2008); Brutti et al. (2011); Dong et al. (2012); Cai et al. (2014); Stallard (1998); Stallard et al. (1999); Stallard (2003); Jung et al. (2004); Zhao and Woodworth (2009); Zhao et al. (2012); Mander et al. (2012).

Single-arm adaptive group-sequential designs:
Green and Dahlberg (1992); Chen and Ng (1998); Lin and Shih (2004); Banerjee and Tsiatis (2006); Wu and Liu (2007); Jones and Holmgren (2007); Masaki et al. (2009); Tournoux-Facon et al. (2011); Roberts and Ramakrishnan (2011); Jin and Wei (2012); Wunder et al. (2012); Englert and Kieser (2012a); (Englert and Kieser, 2012b); Englert and Kieser (2013); Sambucini (2010); Banerjee and Tsiatis (2006); Chen and Smith (2009).

Table A.2: Comparative designs

Comparative single-stage designs:
Herson and Carter (1986); Thall and Simon (1990); Chang et al. (1999); Hong and Wang (2007); Mayo et al. (2010); Hou et al. (2013); Jung and Sargent (2014).

Comparative classical group-sequential designs:
Chang et al. (1999); Hong and Wang (2007); Jung and Sargent (2014); Jung (2008); Sun et al. (2009); Whitehead et al. (2009); Zhang et al. (2011); Wilding et al. (2012); An et al. (2012); Wason and Mander (2012); Shan et al. (2013); Carsten and Chen (2015); Cronin et al. (1999); Huang et al. (2009); Cellamare and Sambucini (2015).

Comparative adaptive group-sequential designs:
Song (2014); Bersimis et al. (2015); Zhong et al. (2013).

Table A.3: Screening designs

Screening single-stage designs:
Whitehead (1985).

Screening classical group-sequential designs:
Thall et al. (1988, 1989); Cheung (2009); Wason and Jaki (2012); Estey and Thall (2003); Ding et al. (2008); Hee and Stallard (2012).

Screening adaptive group-sequential designs:
Logan (2005); Su (2010); Fan et al. (2011).

Table A.4: Master protocol designs

Basket trials:
Leblanc et al. (2009); Cunanan et al. (2017); Liu et al. (2017); Neuenschwander et al. (2016); Thall et al. (2003); Magnusson and Turnbull (2013); Simon et al. (2016); Yuan et al. (2016); Heinrich et al. (2008); Hyman et al. (2015).

Umrella trials:
Renfro and Sargent (2017); Barroilhet and Matulonis (2018); Kim et al. (2011); Ferrarotto et al. (2015); Barker et al. (2009).

Platform trials:
Berry (2015); Hobbs et al. (2018); Kaplan (2015); Lin and Bunn (2017); Saville and Berry (2016).

Appendix B

Estimation methods in non-oncology phase II designs – a brief literature review

In areas other than oncology phase II, various estimation methods have been proposed for different group-sequential and adaptive designs. In group-sequential parallel designs, common in psychiatry, in which the initial parallel trial with placebo versus experimental drug is augmented by a second parallel trial of placebo versus drug in the placebo non-responders from the initial trial, Tamura et al. (2011) proposed the constrained maximum likelihood estimator (MLE), linear combination estimator and allocation weighted estimator for binary endpoint. For a two-stage continuous endpoint diagnostic biomarker design allowing stopping for futility, Koopmeiners et al. (2012) proposed the unadjusted conditional estimator (sample mean conditional on study completion), bias-corrected conditional estimator, and conditional estimation of receiver operating characteristic (ROC) and positive and negative predictive value curves. Various authors discussed estimators for continuous endpoint multi-stage designs having random sample sizes and deterministic and random stopping rules. These include the conditional MLE (Milanzi et al., 2015, 2014; Molenberghs et al., 2014), mean unbiased and bias-adjusted estimators (Milanzi et al., 2015; Todd et al., 1996), and Rao's bias-adjusted estimator (Emerson and Fleming, 1990; Milanzi et al., 2015). Liu et al. (2006) proposed Rao-Blackwell unbiased and truncation-adaptable unbiased estimator for multi-stage designs in which the endpoint follows a distribution in a one-parameter exponential family. Marginal, overall and unconditional bias-corrected estimators (Emerson and Fleming, 1990; Pinheiro and DeMets, 1997; Whitehead, 1986), and conditional bias-corrected estimators (Fan et al., 2004) have been proposed for multi-stage designs in which the sequential test statistics can be approximated by a Brownian motion with a drift parameter. Shen (2001) and

Stallard and Todd (2005) proposed bias-corrected estimators and Stallard et al. (2008) proposed the approximately conditionally unbiased estimator for continuous endpoint single-stage trials comparing two experimental treatments to select the one with largest mean. Luo et al. (2010) proposed an estimator based on conditional moments for binary endpoint two-stage designs with adaptive treatment selection (drop-the-losers designs). For drop-the-losers designs with continuous endpoint, several estimation approaches have also been proposed, including the shrinkage estimation (Bowden et al., 2014; Carreras and Brannath, 2013; Carter and Rolph, 1974), overall mean (Posch et al., 2005), conditional conditional MLE (Bebu et al., 2010), uniformly minimum variance conditionally unbiased estimator (Cohen and Sackrowitz, 1989), extended uniformly minimum variance conditionally unbiased estimator (Bowden and Glimm, 2008), and estimation based on marked point process and stochastic calculus (Luo et al., 2012). For drop-the-losers designs with more than two-stages, Stallard and Todd (2005) proposed the bias-adjusted estimator, and Bowden and Glimm (2014) proposed conditionally unbiased and near unbiased estimators. Brannath et al. (2006) discussed point and interval estimation in flexible two-stage designs with continuous endpoint. In binary endpoint multi-stage designs with response adaptive randomization, Bowden and Trippa (2015) proposed simple bias-corrected and *Rao-Blackwellization* estimators. Coburger and Wassmer (2003) and Cheng and Shen (2004) discussed estimation following a continuous endpoint multi-stage parallel design trial comparing treatment and control, with sample size re-estimation. Median unbiased estimator was proposed by Gao et al. (2013) on similar designs allowing for other endpoints and adaptations.

Appendix C

R code – R package documentation

Package ‘InferenceBEAGSD’

August 23, 2018

Type Package

Title Inference for Binary Endpoint Adaptive Group-Sequential Designs

Version 0.1.0

Author Arsenio Nhacolo

Maintainer Arsenio Nhacolo <anhacolo@uni-bremen.de>

Description This package implents the inference methods proposed in the doctoral thesis intitled 'Bias and precision in early phase adaptive oncology studies and its consequences for confirmatory trials' authored by Arsenio Nhacolo. It includes functions for comparing the performance of various estimators for classical two-stage group-sequential designs with binary endpoint popular in oncology Phase II clinical trials, new inference methods (p-values, and point and interval estimates) proposed for adaptive versions of these designs, and new methods for estimating adjustment factors in order to get an adequately powered Phase III trials when planned based on Phase II trial data.

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 6.1.0

R topics documented:

adjustMet1 *Phase II efficacy estimates/Phase III sample size adjustment factors (Method 1).*

Description

adjustMet1 calculates the multiplicative adjustment factor f to be applied to Phase II efficacy estimate, and the factor ρ to be applied to Phase III sample size estimate using Method 1 proposed by Nhacolo and Brannath (submitted, 2018).

Usage

```
adjustMet1(p2d, p2r, p2e, p2p0 = NULL, p2p1 = NULL, p2a = NULL,
  p2b = NULL, p3p0 = NULL, p3p1 = NULL, p3a = NULL, p3b = NULL,
  nsimul = 5000, seed = NULL)
```

Arguments

p2d Dataframe with Phase II design, with similar as in EKOptAdaptDesigns.

p2r Dataframe containing results of Phase II trials following the design p2d. It is the output of the function AnalyzeEKOAD.

p2e Phase II estimate to consider among the estimates used by codeAnalyzeEKOAD. It can be "pip" (naive MLE) or one of the four estimates from methods proposed by Nhacolo and Brannath (2018): "pim1", "pim2", "pim2v2" or "pim3".

p2p0 Phase II response rate under H_0. If NULL (default), the value is taken p2d.

p2p1 Phase II response rate under H_1. If NULL (default), the value is taken p2d.

p2a Phase II type I error rate. If NULL (default), the value is taken p2d.

p2b Phase II type II error rate. If NULL (default), the value is taken p2d.

p3p0 Phase III response rate of the control group. If NULL (default), the value is set to p2p0.

p3p1 Phase III response rate of the treatment group. If NULL (default), the value is set to p2p1.

p3a Phase III type I error rate. If NULL (default), the value is set p2a.

p3b hase III type II error rate. If NULL (default), the value is set p2b.

nsimul Number of (parametric) bootstrap samples (default 5000).

seed Seed for random number generator. If NULL (default), no seed is set.

Details

The aim of the adjustment is to get an adequately powered Phase III trial based on Phase II data. See the documentation of the function AnIItoIIIRe for more details about the designs.

Value

A list containing two dataframes final and intermed. final contains the final measures for the adjustment factors (f and ρ) and power. intermed holds the intermediate results (of each bootstrap sample).

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Using Estimates from Adaptive Phase II Oncology Trials to Plan Phase III Trials. *Manuscript submitted for publication*, 2018.

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

Ahn, C., Heo, M. and Zhang, S. *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. CRC Press, 2014.

See Also

adjustMet1, SimulateEKOAD, AnalyzeEKOAD.

Examples

```
## Not run:
vdid <- c(6,10) # design ids
vp2est <- c(”pip”,”pim1”,”pim2”,”pim2v2”,”pim3”)
nse <- 1000#number of simulations for each phase
cur <- 1; tot <- length(vdid)*length(vp2est)
for (did in vdid){
 for (p2est in vp2est){
   cat('Processing ',cur,' of ',tot,' (',100*round(cur/tot,1),'%)\n',sep = '')
    load(paste0(”p2r”,did,”.rdata”)) # output of the function AnalyzeEKOAD
   out <- adjustMet1(p2d = EKOADwn[EKOADwn$id==did,], p2r = rslt[1:nse,], p2e = p2est, nsimul = nse, seed = 334
    write.csv(out$final,file = paste0(”final”,did,p2est,”.csv”),row.names = FALSE)
   write.csv(out$intermed,file = paste0(”intermed”,did,p2est,”.csv”),row.names = FALSE)
   cur <- cur+1
 }
}


vdid <- c(6,10)
vp2est <- c(”pip”,”pim1”,”pim2”,”pim2v2”,”pim3”)
fa <- data.frame()
for (did in vdid)
{
 for (p2est in vp2est){
    f <- read.csv(paste0(”final”,did,p2est,”.csv”))
    fn <- names(f)
    f$dsgn <- did
    f <- f[,c('dsgn',fn)]
    fa <- rbind(fa,f)
  }
}
write.csv(fa,file = ”final_all.csv”,row.names = FALSE)

## End(Not run)
```

---

adjustMet2 *Phase III sample size adjustment factor (Method 2).*

Description

adjustMet2 calculates the multiplicative adjustment factor ρ to be applied to Phase III sample size estimate using Method 2 proposed by Nhacolo and Brannath (submitted, 2018).

Usage

```
adjustMet2(p2d, p2r, p2e, p2p0 = NULL, p2p1 = NULL, p2a = NULL,
  p2b = NULL, p3p0 = NULL, p3p1 = NULL, p3a = NULL, p3b = NULL,
  nsimul = 5000, seed = NULL, rhorange = c(0.5, 5), p3mpt = 0.001,
  rhot = 1e-04)
```

Arguments

p2d Dataframe with Phase II design, with similar as in EKOptAdaptDesigns.

`p2r` Dataframe containing results of Phase II trials following the design `p2d`. It is the output of the function `AnalyzeEKOAD`.

`p2e` Phase II estimate to consider among the estimates used by codeAnalyzeEKOAD. It can be `"pip"` (naive MLE) or one of the four estimates from methods proposed by Nhacolo and Brannath (2018): `"pim1"`, `"pim2"`, `"pim2v2"` or `"pim3"`.

`p2p0` Phase II response rate under H_0. If `NULL` (default), the value is taken `p2d`.

`p2p1` Phase II response rate under H_1. If `NULL` (default), the value is taken `p2d`.

`p2a` Phase II type I error rate. If `NULL` (default), the value is taken `p2d`.

`p2b` Phase II type II error rate. If `NULL` (default), the value is taken `p2d`.

`p3p0` Phase III response rate of the control group. If `NULL` (default), the value is set to `p2p0`.

`p3p1` Phase III response rate of the treatment group. If `NULL` (default), the value is set to `p2p1`.

`p3a` Phase III type I error rate. If `NULL` (default), the value is set `p2a`.

`p3b` hase III type II error rate. If `NULL` (default), the value is set `p2b`.

`nsimul` Number of (parametric) bootstrap samples (default 5000).

`seed` Seed for random number generator. If `NULL` (default), no seed is set.

`rhorange` A vector specifying a range to search for ρ. The default is `c(0.5,5)`.

`p3mpt` Tolerated error margin for the power, i.e., maximum allowed absolute difference between the estimated expected power and the target. The default is 0.001.

`rhot` Search for ρ is interrupted and deem unsuccessful if the absolute difference between current and the previous is less than or equal to `rhot`.

Details

The aim of the adjustment is to get an adequately powered Phase III trial based on Phase II data. ρ is found using numerical search. See the documentation of the function `AnIItoIIIRe` for more details about the designs.

Value

A list containing two dataframes `final` and `intermed`. `final` contains the final measures for the adjustment factor (ρ), and for the unadjusted and adjusted power. `intermed` holds the intermediate results (of each bootstrap sample).

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Using Estimates from Adaptive Phase II Oncology Trials to Plan Phase III Trials. *Manuscript submitted for publication*, 2018.

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

Ahn, C., Heo, M. and Zhang, S. *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. CRC Press, 2014.

See Also

adjustMet2, SimulateEKOAD, AnalyzeEKOAD.

Examples

```
## Not run:
vdid <- c(6,10) # design ids
vp2est <- c("pip","pim1","pim2","pim2v2","pim3")
nse <- 1000#number of simulations for each phase
cur <- 1; tot <- length(vdid)*length(vp2est)
for (did in vdid){
 for (p2est in vp2est){
   cat('Processing ',cur,' of ',tot,' (',100*round(cur/tot,1),'%)\n',sep = '')
    load(paste0("p2r",did,".rdata")) # output of the function AnalyzeEKOAD
   out <- adjustMet2(p2d = EKOADwn[EKOADwn$id==did,], p2r = rslt[1:nse,], p2e = p2est, nsimul = nse, seed = 334
    write.csv(out$final,file = paste0("final",did,p2est,".csv"),row.names = FALSE)
   write.csv(out$intermed,file = paste0("intermed",did,p2est,".csv"),row.names = FALSE)
   cur <- cur+1
 }
}


vdid <- c(6,10)
vp2est <- c("pip","pim1","pim2","pim2v2","pim3")
fa <- data.frame()
for (did in vdid)
{
 for (p2est in vp2est){
    f <- read.csv(paste0("final",did,p2est,".csv"))
    fn <- names(f)
    f$dsgn <- did
    f <- f[,c('dsgn',fn)]
    fa <- rbind(fa,f)
  }
}
write.csv(fa,file = "final_all.csv",row.names = FALSE)

## End(Not run)
```

AnalyzeEKOAD *Analyse simulated adaptive trials.*

Description

AnalyzeEKOAD performs inference on trials simulated by the function SimulateEKOAD using the methods proposed by Nhacolo and Brannath (2018) and naive maximum likelihood.

Usage

```
AnalyzeEKOAD(replicates = NULL, basedir = NULL)
```

Arguments

`replicates`	Number of simulated trials to be analysed. If `NULL` (default), all trials found in `./basedir/SimulatedTrials` are analysed.
`basedir`	The base directory containing the sub-directory `SimulatedTrials` with the simulated trials. If `NULL` (default), the current working directory is uded.

Details

Overall p-values, point estimates and confidence intervals are calculated.

Value

A dataframe with the results. A copy is saved in the file `Results.csv` in the `basedir`.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

`SimulateEKOAD`, `mue1`, `mue2`, `mue2v2`, `mue3`.

`AnalyzePerformanceSimon`

Performance of estimation methods

Description

It takes the results produced by `AnalyzeSimonDsgn` and `AnalyzeSimonDsgnAdaptN` and produces a dataframe containing bias, mean square error and variance of the estimators. It also calculates the power and the expected sample size (EN) where applicable.

Usage

```
AnalyzePerformanceSimon(designs = "all", basedir = NA)
```

Arguments

`designs`	Taking values `"fixed"`, `"adaptive"` or `"all"`, indicating whether only classical, adaptive or all designs should be included. The default is `"all"`.
`basedir`	The root directory in which simulations were performed. The current working directory is assumed by default. It must contain all the files and folders created by `SimulateSimonDsgn` and/or `SimulateSimonDsgnAdaptN`.

Details

Computations are done for different combinations of values of `stop`, (0,1), and `success`, (0,1). See `AnalyzeSimonDsgn` or `AnalyzeSimonDsgnAdaptN`. For instance, computations done on all simulated trials are marked with `"both"` in the columns `stop` and `success`, while the ones done only on trials that continued to the final stage have `stop = "no"` and `success = "both"`.

Value

Dataframe containing bias, mean square error and variance of the estimators, power, expected sample size, and design information.

Author(s)

Arsenio Nhacolo

See Also

`AnalyzeSimonDsgn`, `AnalyzeSimonDsgnAdaptN`, `pdata` and `AnalyzePerformanceSimon2`.

Examples

```
## Not run:
AnalyzePerformanceSimon()

## End(Not run)
```

`AnalyzePerformanceSimon2` *Performance of estimation methods*

Description

It takes the results produced by `AnalyzeSimonDsgn` and `AnalyzeSimonDsgnAdaptN` and produces a dataframe containing bias, mean square error and variance of the estimators. It also calculates the power and the expected sample size (EN) where applicable.

Usage

```
AnalyzePerformanceSimon2(designs = "all", basedir = NA)
```

Arguments

`designs`	Taking values `"fixed"`, `"adaptive"` or `"all"`, indicating whether only classical, adaptive or all designs should be included. The default is `"all"`.
`basedir`	The root directory in which simulations were performed. The current working directory is assumed by default. It must contain all the files and folders created by `SimulateSimonDsgn` and/or `SimulateSimonDsgnAdaptN`.

Details

It is the same as `AnalyzePerformanceSimon`, but here the estimation is done only for two sets: all trials (unconditional), and only trials that continued to final stage (conditional).

Value

Dataframe containing bias, mean square error and variance of the estimators, power, expected sample size, and design information.

Author(s)

Arsenio Nhacolo

See Also

AnalyzeSimonDsgn, AnalyzeSimonDsgnAdaptN, pdata and AnalyzePerformanceSimon.

Examples

```
## Not run:
AnalyzePerformanceSimon2()

# Simulation example
seed = 1986
p0 <- 0.1
alpha <- 0.05
beta <- 0.1
repl <- 100 # number of replicated trials for each p
if (file.exists("PerforAll.csv")) unlink("PerforAll.csv")
coln <- TRUE
while (p0 < 0.5){
  pv <- seq(p0+0.2,p0+0.4,0.1) # p to simulate data
  p1v <- seq(p0+0.2,p0+0.3,0.1) # p to get design
  for (p1 in p1v){
    designParam <- CalculateSimonDsgn(p0, p1, alpha, beta)
    pstart <- p0+0.1
    SimulateSimonDsgn(repl, designParam, pstart, seed = seed)
    SimulateSimonDsgnAdaptN(repl, designParam, pstart, seed = seed)
    AnalyzeSimonDsgn()
    AnalyzeSimonDsgnAdaptN()
    perf <- AnalyzePerformanceSimon2()
    for (p in pv){
      SimulateSimonDsgn(repl, designParam, p, seed = seed)
      SimulateSimonDsgnAdaptN(repl, designParam, p, seed = seed)
      AnalyzeSimonDsgn()
      AnalyzeSimonDsgnAdaptN()
      perf <- rbind(perf, AnalyzePerformanceSimon2())
    }
    write.csv(perf, file = paste("PerforAll_a",alpha,"b",beta,"p0",p0,"p1",
                             p1,".csv", sep = ""), row.names = F)
   write.table(perf, file ="PerforAll.csv", append = T, sep = ",", row.names = F, col.names = coln)
    coln <- FALSE
  }
  p0 <- p0+0.1
}

## End(Not run)
```

AnalyzeSimonDsgn *Analysis of simulated Simon's design trials*

Description

Analyses the trials simulated by SimulateSimonDsgn.

Usage

```
AnalyzeSimonDsgn(replicates = NA, basedir = NA)
```

Arguments

replicates	Number of trials to be analysed. By default all simulated trials are analysed.
basedir	The root directory in which simulations were performed. The current working directory is assumed by default. It must contain all the files and folders created by SimulateSimonDsgn.

Details

In addition to hypothesis testing, the response rate is estimated using different estimators: pm, pg, pu, pp and pk.

Value

Creates two data files in basedir containing results for optimal (*ResultsOptimalDesign.csv*) and minimax (*ResultsMinimaxDesign.csv*). The files contain a trial ID, stage 1, stage 2 and overall number of successful responses, s1, s2 and s, sample sizes (equal to those pre-specified by design), n1, n2 and n, and critical values, r1 and r. p0 the response rate assumed under H_0 and dsgnp1 under H_1. p1 is the true response rate (used for generating trial data). pm1 and pm2 are, respectively, pm based only of stage 1 and stage 2 data. stop indicates whether the trial stopped at first stage (stop = 1), and success indicates whether H_0 was rejected (success = 1).

Author(s)

Arsenio Nhacolo

See Also

CalculateSimonDsgn, SimulateSimonDsgn, AnalyzePerformanceSimon and AnalyzeSimonDsgnAdaptN.

Examples

```
AnalyzeSimonDsgn()
```

AnalyzeSimonDsgnAdaptN *Analysis of simulated adaptive Simon's design trials*

Description

Analyses the trials simulated by SimulateSimonDsgnAdaptN.

Usage

```
AnalyzeSimonDsgnAdaptN(replicates = NA, basedir = NA)
```

Arguments

replicates	Number of trials to be analysed. By default all simulated trials are analysed.
basedir	The root directory in which simulations were performed. The current working directory is assumed by default. It must contain all the files and folders created by SimulateSimonDsgnAdaptN.

Details

In addition to hypothesis testing, the response rate is estimated using different estimators: pm, pg, pu, pp and pk. The overall critical value, r, is recalculated using conditional type I error (*Englert and Kieser, 2012*).

Value

Creates two data files in basedir containing results for optimal (*ResultsOptimalDesignAdapt.csv*) and minimax (*ResultsMinimaxDesignAdapt.csv*). The files contain a trial ID, stage 1, stage 2 and overall number of successful responses, s1, s2 and s, sample sizes (equal to those pre-specified by design), n1, n2 and n, and critical values, r1 and r. p0 the response rate assumed under H_0 and dsgnp1 under H_1. p1 is the true response rate (used for generating trial data). pm1 and pm2 are, respectively, pm based only of stage 1 and stage 2 data. stop indicates whether the trial stopped at first stage (stop = 1), and success indicates whether H_0 was rejected (success = 1).

Author(s)

Arsenio Nhacolo

See Also

CalculateSimonDsgn, SimulateSimonDsgnAdaptN, AnalyzePerformanceSimon and AnalyzeSimonDsgn.

Examples

```
AnalyzeSimonDsgnAdaptN()
```

`AnIItoIIIRe` *Use of Phase II estimates to plan Phase III sample size.*

Description

`AnIItoIIIRe` calculates the power in a Phase III equal-size group two-arm randomized clinical trial with a binary response planned using estimates from Phase II adaptive two-stage trial.

Usage

```
AnIItoIIIRe(rslt, f = c(0.95, 0.96, 0.97, 0.98, 0.99))
```

Arguments

`rslt`	Dataframe containing the output from the function `AnalyzeEKOAD`, but with only successful trials (rslt$suco==1), i.e., trials in which H_0 was rejected.
`f`	Vector of length 5 containing multiplicative adjustment factors to be applied to Phase II estimates. The default is `f = c(.95,.96,.97,.98,.99)`.

Details

The sample size (N) of the Phase III trial is based on the estimates naive MLE and estimators proposed by Nhacolo and Brannath (2018). Different values of retention factor f proposed by Kirby et al. (2012) are applied. The control group response rate is considered to be equal to that under the null hypothesis of the Phase II design, and the hypothesized treatment group response rate considered to be equal to that estimated from the Phase II trial. The target type I error and power are the same as of the Phase II design. Two-sided hypothesis test is assumed is a sample size per group, and equal size groups are assume. Hence, N total is 2*N. When calculating the power, the true response rate (in treatment group) is considered to be the one under which the Phase II trial was simulated (spi1).

Value

The input dataframe with corresponding Phase III sample size and power.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Using Estimates from Adaptive Phase II Oncology Trials to Plan Phase III Trials. *Manuscript submitted for publication*, 2018.

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

Ahn, C., Heo, M. and Zhang, S. *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. CRC Press, 2014.

See Also

`AnalyzeEKOAD`, `SimulateEKOAD`, `PerforIItoIIIRe`.

aop1 *Overall p-value (Method 1 of Nhacolo and Brannath, 2018).*

Description

aop1 calculates the overall p-value for adaptive two-stage designs with binary endpoint using the Method 1 (see Nhacolo and Brannath, 2018).

Usage

```
aop1(dsgn, x1o, xo, verbose = TRUE)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.
verbose	If TRUE (default) messages will be printed.

Details

This is one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

p-value.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

aop2, aop2v2, aop3e.

aop1e *Overall p-value for CI (Method 1 of Nhacolo and Brannath, 2018).*

Description

aop1e is a modified version of aop1 used for getting the confidence interval.

Usage

```
aop1e(dsgn, x1o, xo, newpi0)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.
newpi0	New response probability that replaces the one under the null hypothesis.

Details

This is one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

p-value.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

aop2e, aop2ev2, aop3e, aop1.

aop2 *Overall p-value (Method 2 of Nhacolo and Brannath, 2018).*

Description

aop2 calculates the overall p-value for adaptive two-stage designs with binary endpoint using the Method 2 (see Nhacolo and Brannath, 2018).

Usage

```
aop2(dsgn, x1o, xo, verbose = TRUE)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.
verbose	If TRUE (default) messages will be printed.

Details

This is one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

p-value.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

aop1, aop2v2, aop3e

aop2e *Overall p-value for CI (Method 2 of Nhacolo and Brannath, 2018).*

Description

aop2e is a modified version of aop2 used for getting the confidence interval.

Usage

```
aop2e(dsgn, x1o, xo, newpi0)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.
newpi0	New response probability that replaces the one under the null hypothesis.

Details

This is one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

p-value.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

aop1e, aop2ev2, aop3e, aop2.

aop2ev2 *Overall p-value for CI (Method 2v2 of Nhacolo and Brannath, 2018).*

Description

aop2ev2 is a modified version of aop2v2 used for getting the confidence interval.

Usage

```
aop2ev2(dsgn, x1o, xo, newpi0)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.
newpi0	New response probability that replaces the one under the null hypothesis.

Details

This is one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

p-value.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

aop1e, aop2e, aop3e, aop2v2.

aop2v2 *Overall p-value (Method 2v2 of Nhacolo and Brannath, 2018).*

Description

aop2v2 calculates the overall p-value for adaptive two-stage designs with binary endpoint using the Method 2v2 (see Nhacolo and Brannath, 2018).

Usage

```
aop2v2(dsgn, x1o, xo, verbose = TRUE)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.
verbose	If TRUE (default) messages will be printed.

Details

This is one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

p-value.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

aop1, aop2

`aop3e` *Overall p-value (Method 3 of Nhacolo and Brannath, 2018).*

Description

`aop3e` calculates the overall p-value for adaptive two-stage designs with binary endpoint using the Method 3 (see Nhacolo and Brannath, 2018).

Usage

```
aop3e(dsgn, x1o, xo, newpi0 = NULL)
```

Arguments

`dsgn`	Dataframe containing one of the designs in `EKOADwn`.
`x1o`	The observed stage 1 number of responses.
`xo`	The total observed number of responses.
`newpi0`	New response probability that replaces the one under the null hypothesis. Omit it if the intention is only to calculate the overall p-value.

Details

This is one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

p-value.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

`aop1`, `aop1e`, `aop2`, `aop2e`, `aop2v2`, `aop2ev2`

`CalculateSimonDsgn` *Simon's designs*

Description

`CalculateSimonDsgn` finds Simon's optimal and minimax designs.

Usage

```
CalculateSimonDsgn(p0, p1, alpha, beta, verbose = TRUE)
```

Arguments

`p0`	The response rate under the null hypothesis.
`p1`	The response rate under the alternative hypothesis.
`alpha`	Type I error rate.
`beta`	Type II error rate.
`verbose`	If TRUE (default) the designs are printed (gives messy printout when the function is run without assignment).

Details

Simon's designs are two-stage single-arm for phase II clinical trials. They consist in first stage and overall sample sizes and critical values, `n1` and `n`, and `r1` and `r`, respectively.

Value

A two-row dataframe containing the optimal and the minimax designs.

Author(s)

Arsenio Nhacolo

References

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

See Also

`SimulateSimonDsgn` and `SimulateSimonDsgnAdaptN`.

Examples

```
d <- CalculateSimonDsgn(0.2, 0.4, 0.05, 0.1)
```

`checkMonoDCF` *Check the monotonicity of the sample space ordering.*

Description

`checkMonoDCF` checks the monotonicity of the sample space ordering defined based on inverse normal combination function (see Nhacolo and Brannath, 2018).

Usage

```
checkMonoDCF(d, verbose = TRUE)
```

Arguments

`d`	Dataframe containing one of the designs in `EKOADwn`.
`verbose`	If `TRUE` (default) messages about monotonicity will be printed.

Details

The monotonicity is with respect to the stage 2 number of successes.

Value

A list containing a dataframe (`mono`) with detailed info, and a logical variable `notmono` indicating whether non-monotonicity was concluded.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

Examples

```
## Not run:
#Check for all Englert and Kieser designs
notmov <- c()
for (i in 1:max(EKOADwn$id)){
  notmov <- c(notmov,checkMonoDCF(EKOADwn[EKOADwn$id==1,],verbose=FALSE)[[2]])
}
isMonotone <- !any(notmov);isMonotone

## End(Not run)
```

ci1 *Confidence interval (using Method 1 of Nhacolo and Brannath, 2018).*

Description

ci1 computes confidence interval.

Usage

```
ci1(dsgn, x1o, xo, alpha = 0.05, twosided = FALSE)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.
alpha	The significance level.
twosided	If FALSE (default) a one-sided CI is produced.

Details

This CI is obtained using the Method 1, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

CI is a list with lower and upper bounds.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

ci2, ci2v2, ci3, aop1, aop1e, pipv1, , mue1.

`ci2` *Confidence interval (using Method 2 of Nhacolo and Brannath, 2018).*

Description

`ci2` computes confidence interval.

Usage

```
ci2(dsgn, x1o, xo, alpha = 0.05, twosided = FALSE)
```

Arguments

`dsgn`	Dataframe containing one of the designs in `EKOADwn`.
`x1o`	The observed stage 1 number of responses.
`xo`	The total observed number of responses.
`alpha`	The significance level.
`twosided`	If `FALSE` (default) a one-sided CI is produced.

Details

This CI is obtained using the Method 2, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

CI is a list with lower and upper bounds.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

`ci1`, `ci2v2`, `ci3`, `aop2`, `aop2e`, `pipv2`, `mue2`.

ci2v2 *Confidence interval (using Method 2v2 of Nhacolo and Brannath, 2018).*

Description

ci2v2 computes confidence interval.

Usage

```
ci2v2(dsgn, x1o, xo, alpha = 0.05, twosided = FALSE)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.
alpha	The significance level.
twosided	If FALSE (default) a one-sided CI is produced.

Details

This CI is obtained using the Method 2v2, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

CI is a list with lower and upper bounds.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

ci1, ci2, ci3, aop2v2, aop2ev2, pipv2v2, , mue2v2.

`ci3` *Confidence interval (using Method 3 of Nhacolo and Brannath, 2018).*

Description

`ci3` computes confidence interval.

Usage

```
ci3(dsgn, x1o, xo, alpha = 0.05, twosided = FALSE)
```

Arguments

`dsgn`	Dataframe containing one of the designs in `EKOADwn`.
`x1o`	The observed stage 1 number of responses.
`xo`	The total observed number of responses.
`alpha`	The significance level.
`twosided`	If `FALSE` (default) a one-sided CI is produced.

Details

This CI is obtained using the Method 3, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

CI is a list with lower and upper bounds.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

`ci1`, `ci2`, `ci2v2`, `aop3e`, `pipv3`, `mue3`.

`dsgnPrep` *Pre-process the Englert and Kieser (2013) optimal adaptive designs.*

Description

`dsgnPrep` takes Englert and Kieser's optimal adaptive design and adds information that is needed by other functions.

Usage

```
dsgnPrep(dsgn = NULL, w1 = "n", w2 = NULL)
```

Arguments

`dsgn` Dataframe containing one of the designs in `EKOptAdaptDesigns`.

`w1, w2` Stage 1 and 2 weights. If `w1="n"` (default), weights a calculated based on stage-wise sample sizes as described in Nhacolo and Brannath (2018). If `w1="sr2"`, then `w1=w2=1/sqrt(2)`.

Details

The function adds, to each x1 leading to 2nd stage, the corresponding p-value (p1) and its backwards image (p1B), the stage-wise weights w1 and w2 and other information used in inference methods proposed by Nhacolo and Brannath (2018).

Value

Dataframe containing the input dataframe with added information.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

Examples

```
## Not run:
#Designs with w1a and w2 calculated based on sample sizes
EKOADwn <- data.frame()
for (j in 1:max(EKOptAdaptDesigns$id)){
 EKOADwn <- rbind(EKOADwn, dsgnPrep(dsgn = EKOptAdaptDesigns[EKOptAdaptDesigns$id==j,],w1 = "n"))
}
save(EKOADwn,file = "EKOADwn.RData")

## End(Not run)
```

EKOADwn *Pre-processed Englert and Kieser (2013)'s optimal adaptive designs.*

Description

A dataframe containing all the designs in EKOptAdaptDesigns pre-processed by the function dsgnPrep, the argument w1 set to "n".

Usage

```
EKOADwn
```

Format

A dataframe with 709 rows and 20 variables.

EKOptAdaptDesigns *Englert and Kieser (2013)'s optimal adaptive designs.*

Description

A dataframe containing all optimal adaptive two-stage designs for phase II cancer clinical trials present in Englert and Kieser (2013).

Usage

```
EKOptAdaptDesigns
```

Format

A dataframe with 709 rows and 11 variables:

id Identifier of the designs

x1 Number of successes (responses) at stage 1

n2 Stage 2 sample size

D Discrete conditional error function

l Stage 2 decision boundary

pi0 Response probability under the null hypothesis

pi1 Response probability under the alternative hypothesis

alpha Type I error rate

beta Type II error rate

n1 Stage 1 sample size

n2max Maximum stage 2 sample size

Source

https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201200220 getN2v2 *Number of patients to be enrolled in the second stage*

Description

Calculates the number of patients which should be enrolled in the second stage if the conditional power should be altert to "cp". It's a version of getN2.

Usage

```
getN2v2(cp, p1, design, k, mode = 0, alpha = 0.05)
```

Arguments

cp	conditional power to which the number of patients for the second stage should be adjusted.
p1	response probability under the alternative hypothesis.
design	a dataframe containing all critical values for a Simon's two-stage design defined by the colums r1, n1, r, n and p0. • r1 = critical value for the first stage (more than r1 responses needed to proceed to the second stage). • n1 = number of patients enrolled in the first stage. • r = critical value for the whole trial (more than r responses needed at the end of the study to reject the null hypothesis). • n = number of patients enrolled in the whole trial. • p0 = response probability under the null hypothesis.
k	number of responses observed at the interim analysis.
mode	a value out of 0,1,2,3 dedicating the methode spending the "rest alpha" (difference between nominal alpha level and actual alpha level for the given design). • 0 = "rest alpha" is not used. • 1 = "rest alpha" is spent proportionally. • 2 = "rest alpha" is spent equally. • 3 = "rest alpha" is spent only to the worst case scenario (minimal number of responses at the interim analysis so that the study can proceed to the second stage).
alpha	overall significance level the trial was planned for.

Details

This functon is the same as getN2 (OneArmPhaseTwoStudy package), with some changes in arguments' validation. It's is a helper to SimulateSimonDsgnAdaptN.

References

Englert S., Kieser M. Adaptive designs for single-arm phase II trials in oncology. *Pharm Stat*, 2012, 11, 241-249.

See Also

getN2, SimulateSimonDsgnAdaptN.

Examples

```
designParam <- CalculateSimonDsgn(0.2, 0.4, 0.05, 0.1)
dsgn <- designParam[designParam$Type == "Optimal",]
getN2v2(0.9, dsgn$p1, dsgn, 7)
```

mue1 *Median estimate (using Method 1 of Nhacolo and Brannath, 2018).*

Description

mue1 calculates the median estimate of the response rate.

Usage

```
mue1(dsgn, x1o, xo)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.

Details

This estimate is obtained using the Method 1, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

Median estimate of response probability.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

mue2, mue2v2, mue3, aop1, aop1e, pipv1.

`mue2` *Median estimate (using Method 2 of Nhacolo and Brannath, 2018).*

Description

`mue2` calculates the median estimate of the response rate.

Usage

```
mue2(dsgn, x1o, xo)
```

Arguments

`dsgn`	Dataframe containing one of the designs in `EKOADwn`.
`x1o`	The observed stage 1 number of responses.
`xo`	The total observed number of responses.

Details

This estimate is obtained using the Method 2, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

Median estimate of response probability.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

`mue1`, `mue2v2`, `mue3`, `aop2`, `aop2e`, `pipv2`.

mue2v2 *Median estimate (using Method 2v2 of Nhacolo and Brannath, 2018).*

Description

mue2v2 calculates the median estimate of the response rate.

Usage

```
mue2v2(dsgn, x1o, xo)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.

Details

This estimate is obtained using the Method 2v2, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

Median estimate of response probability.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

mue1, mue2, mue3, aop2v2, aop2ev2, pipv2v2.

`mue3` *Median estimate (using Method 3 of Nhacolo and Brannath, 2018).*

Description

`mue3` calculates the median estimate of the response rate.

Usage

```
mue3(dsgn, x1o, xo)
```

Arguments

`dsgn`	Dataframe containing one of the designs in `EKOADwn`.
`x1o`	The observed stage 1 number of responses.
`xo`	The total observed number of responses.

Details

This estimate is obtained using the Method 3, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

Median estimate of response probability.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

`mue1`, `mue2`, `mue2v2`, `aop3e`, `pipv3`.

`Nct` *Sample size per group for single-stage parallel-group RCT.*

Description

`Nct` calculates sample size for one group in an equal-size group two-arm randomized clinical trial with a binary response.

Usage

```
Nct(pc, pt, alp = 0.05, pow = 0.8)
```

Arguments

`pc`	Response probability in control group.
`pt`	Response probability in treatment group.
`alp`	Significance level (default: 0.05).
`pow`	Power (default: 0.8)

Details

The sample size is for one group (arm), double the number to get the total.

Value

Sample size for one group.

Author(s)

Arsenio Nhacolo

References

Ahn, C., Heo, M. and Zhang, S. *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research.* CRC Press, 2014.

See Also

`Pwr`.

Examples

```
Nct(0.2,0.3,0.05,0.9)
```

`pdata` *Helper function for analysing the performance of estimators*

Description

It takes the results produced by `AnalyzeSimonDsgn` or `AnalyzeSimonDsgnAdaptN` and produces a dataframe containing bias, mean square error and variance of the estimators.

Usage

```
pdata(t, design, stop, success, replicates)
```

Arguments

`t`	Dataframe containing results produced by `AnalyzeSimonDsgn` or `AnalyzeSimonDsgnAdaptN`.
`stop`	Taking value `"yes"`, `"no"` or `"both"`, indicating that only trials that stopped, continued or both were analysed.
`success`	Taking value `"yes"`, `"no"` or `"both"`, indicating that only trials that were successful, unsuccessful or both were analyzed.
`replicates`	Number of trials analysed. It is equal to the number of rows in `t`.

Details

It is a helper function for `AnalyzePerformanceSimon`. It also calculates the power and the expected sample size (EN) where applicable.

Value

Dataframe containing bias, mean square error and variance of the estimators.

Author(s)

Arsenio Nhacolo

See Also

`AnalyzeSimonDsgn`, `AnalyzeSimonDsgnAdaptN` and `AnalyzePerformanceSimon`.

Examples

```
## Not run:
rslt <- read.csv("ResultsOptimalDesign.csv")
nrep <- nrow(rslt)
t <- rslt
presult <- pdata(t, "Optimal", "both", "both", nrep)
t <- rslt[rslt$stop == 0,]
presult <- rbind(presult, pdata(t, "Optimal", "no", "both", nrep))

## End(Not run)
```

`pdata2` *Helper function for analysing the performance of estimators*

Description

`pdata2` is a helper function used by function `PerformanceEKOAD`.

Usage

```
pdata2(t, stop, success, replicates)
```

Details

Not to be used directly.

Value

Dataframe

Author(s)

Arsenio Nhacolo

`PerforIItoIIIRe` *Performance, with respect to Phase III power, of phase II estimates.*

Description

`PerforIItoIIIRe` calculates the mean and median power in a Phase III trials from the output of `AnIItoIIIRe`.

Usage

```
PerforIItoIIIRe(t)
```

Arguments

`t`	Dataframe containing the output from the function `AnIItoIIIRe`.

Value

Dataframe.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Using Estimates from Adaptive Phase II Oncology Trials to Plan Phase III Trials. *Manuscript submitted for publication*, 2018.

See Also

AnIItoIIIRe.

Examples

```
## Not run:
rslt <- read.csv("ResultsAll.csv")
rsltfull <- rslt
rslt <- rslt[rslt$suco==1,]
rslt <- rslt[,c("pi0", "pi1", "spi1", "alpha", "beta", "suco", "pip",
                "pim1", "pim2", "pim2v2", "pim3")]
rslt <- rslt[rslt$spi1>=rslt$pi0+0.1 & rslt$spi1<=rslt$pi1+0.3,]
rslt$spi1f <- factor(rslt$spi1)
cats <- levels(rslt$spi1f)
ncats <- length(cats)
setwd(paste0("C:/Users/arsenio/Documents/PhD/Simulations/Paper2/Reuse/pi01by0.01/50000/",did))
save(ncats,file = "ncats.rdata")
for (i in 1:ncats){
  sr <- rslt[rslt$spi1f==cats[i],]#Single result (result of a specific spi1)
  save(sr,file = paste0("sr",i,".rdata"))
}
load("ncats.rdata")
PerfAll <- data.frame()
for (k in 1:ncats){
 load(paste0("sr",k,".rdata"))
 sre <- AnIItoIIIRe(rslt = sr,f = c(.95,.96,.97,.98,.99))
  PerfAll <- rbind(PerfAll,PerforIItoIIIRe(sre))
  rm(sr)
}
write.csv(PerfAll, file = "PerfAllIItoIII.csv", row.names = F)

## End(Not run)
```

PerformanceEKOAD *Performance of estimation methods*

Description

PerformanceEKOAD calculates performance measures (bias, mean square error, coverage probability) of the estimation methods based on the results produced by AnalyzeEKOAD.

Usage

```
PerformanceEKOAD(basedir = NULL)
```

Arguments

basedir	The base directory containing the file with the results (Results.csv). If NULL (default), the current working directory is uded.

Value

A dataframe with the performance results. A copy is saved in the file the PerformanceResults.csv in the basedir.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

`SimulateEKOAD`, `AnalyzeEKOAD`.

Examples

```
## Not run:
#SIMULATIONS
for (did in c(6,10)){#Design ID
cat("============================ Design ",did," ============================\n")
repl <- 50000 # number of replicated trials for each p
dir.create(as.character(did))
setwd(as.character(did))
design <- EKOADwn[EKOADwn$id==did,]
seed = 3343
if (file.exists("PerforAll.csv")) unlink("PerforAll.csv")
piv <- seq(0,1,0.025) # p to simulate data
resul <- data.frame()
perf <- data.frame()
k <- 0
pl <- length(piv)
for (pi in piv){
  k <- k+1
 cat("_________________________ pi = ",pi, " (",k," of ",pl,") _________________________\n",sep = "")
  SimulateEKOAD(replicates = repl, dsgn = design, newpi1 = pi, seed = seed)
  resul <- rbind(resul, AnalyzeEKOAD())
  perf <- rbind(perf, PerformanceEKOAD())
}
write.table(resul, file ="ResultsAll.csv", sep = ",", row.names = F, col.names = TRUE)
write.table(perf, file ="PerforAll.csv", sep = ",", row.names = F, col.names = TRUE)
cat("Design ID: ", design$id[1], "\nReplicates: ", repl, "\nSeed: ", seed,
    "\nDate last run: ", date(),file = "info.txt", sep = "", append = FALSE)
}

## End(Not run)
```

---

pg *Bias-reduced estimator*

Description

Calculates the bias-reduced estimator of the true response rate as proposed by *Guo and Liu (2005)*.

Usage

```
pg(s, n1, r1, n)
```

Arguments

`s` Total number of successes.

`n1` Stage 1 sample size.

`r1` Stage 1 critical value (trial is stopped at stage 1 if the number of successes is at most `r1`).

`n` Total sample size.

Details

It uses bias subtraction, with bias calculated by `sbias` and response rate estimated by `pm`.

Value

Estimate of the response rate.

Author(s)

Arsenio Nhacolo

References

Guo, H. Y. and Liu, A. A simple and efficient bias-reduced estimator of response probability following a group sequential phase II trial. *J Biopharm Stat*, 2005, 15, 773-781.

See Also

`sbias`, `pm`, `pu`, `pp` and `pk`.

Examples

```
pg(21, 19, 4, 54)
```

`pipv1` *Response rate to attain a specified p-value (using Method 1 of Nhacolo and Brannath, 2018).*

Description

`pipv1` finds the response probability under the null hypotheisis that, given the observed data, would yield a desired overall p-value.

Usage

```
pipv1(dsgn, x1o, xo, pv)
```

Arguments

`dsgn` Dataframe containing one of the designs in `EKOADwn`.

`x1o` The observed stage 1 number of responses.

`xo` The total observed number of responses.

`pv` The desired p-value.

Details

The p-value is obtained using the Method 1, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

Response probability.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

`pipv2`, `pipv2v2`, `pipv3`, `aop1`, `aop1e`.

`pipv2` *Response rate to attain a specified p-value (using Method 2 of Nhacolo and Brannath, 2018).*

Description

`pipv2` finds the response probability under the null hypothesis that, given the observed data, would yield a desired overall p-value.

Usage

```
pipv2(dsgn, x1o, xo, pv)
```

Arguments

`dsgn`	Dataframe containing one of the designs in `EKOADwn`.
`x1o`	The observed stage 1 number of responses.
`xo`	The total observed number of responses.
`pv`	The desired p-value.

Details

The p-value is obtained using the Method 2, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

Response probability.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

`pipv1`, `pipv2v2`, `pipv3`, `aop2`, `aop2e`.

`pipv2v2` *Response rate to attain a specified p-value (using Method 2v2 of Nhacolo and Brannath, 2018).*

Description

`pipv2v2` finds the response probability under the null hypotheisis that, given the observed data, would yield a desired overall p-value.

Usage

```
pipv2v2(dsgn, x1o, xo, pv)
```

Arguments

`dsgn`	Dataframe containing one of the designs in `EKOADwn`.
`x1o`	The observed stage 1 number of responses.
`xo`	The total observed number of responses.
`pv`	The desired p-value.

Details

The p-value is obtained using the Method 2v2, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

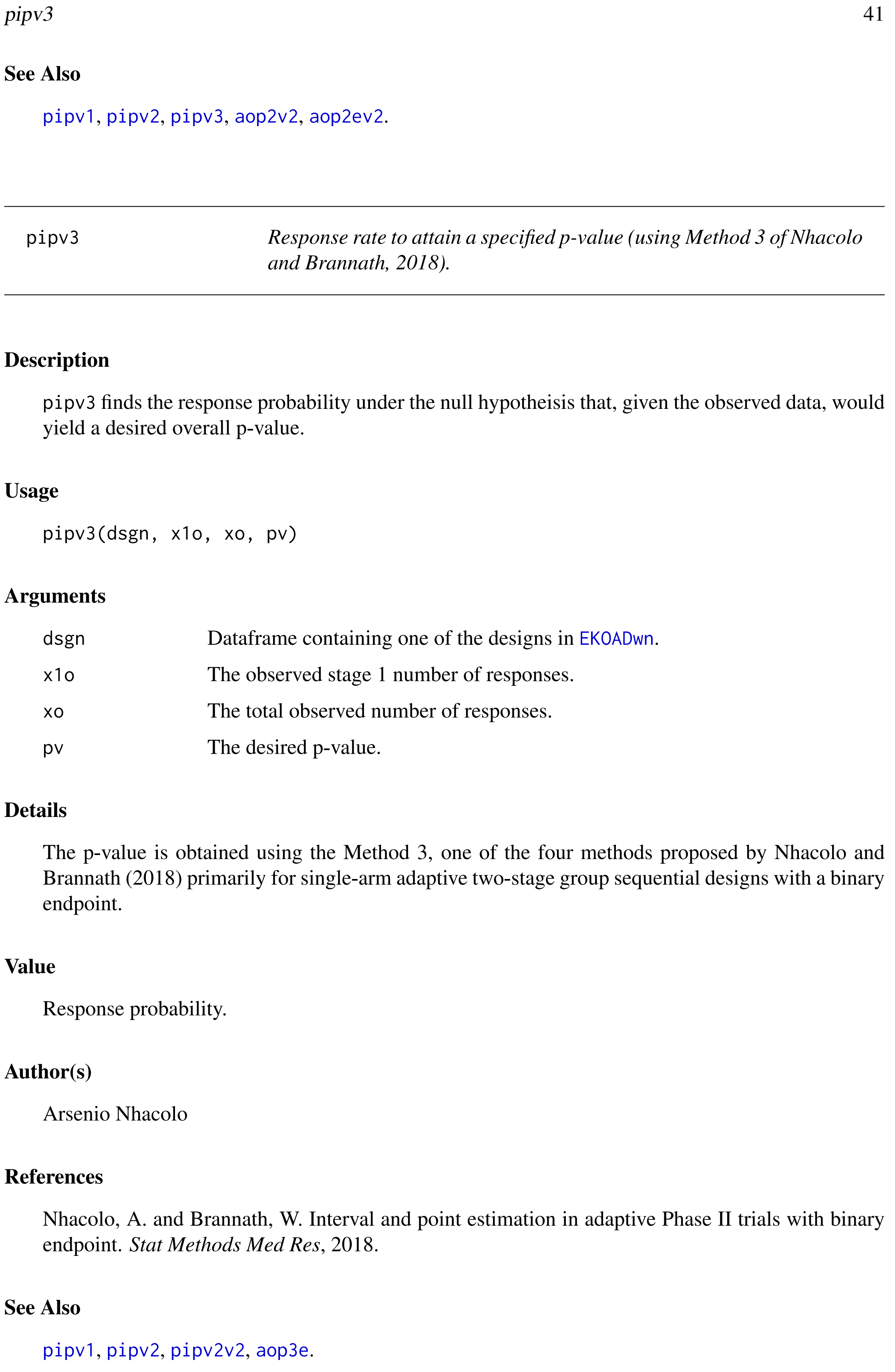
Response probability.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.



See Also

pipv1, pipv2, pipv3, aop2v2, aop2ev2.

pipv3 *Response rate to attain a specified p-value (using Method 3 of Nhacolo and Brannath, 2018).*

Description

pipv3 finds the response probability under the null hypotheisis that, given the observed data, would yield a desired overall p-value.

Usage

```
pipv3(dsgn, x1o, xo, pv)
```

Arguments

dsgn	Dataframe containing one of the designs in EKOADwn.
x1o	The observed stage 1 number of responses.
xo	The total observed number of responses.
pv	The desired p-value.

Details

The p-value is obtained using the Method 3, one of the four methods proposed by Nhacolo and Brannath (2018) primarily for single-arm adaptive two-stage group sequential designs with a binary endpoint.

Value

Response probability.

Author(s)

Arsenio Nhacolo

References

Nhacolo, A. and Brannath, W. Interval and point estimation in adaptive Phase II trials with binary endpoint. *Stat Methods Med Res*, 2018.

See Also

pipv1, pipv2, pipv2v2, aop3e.

pk *Median unbiased estimator*

Description

Calculates the median unbiased estimator of true response rate for Simon-like designs.

Usage

```
pk(s, n1, r1, n, p0)
```

Arguments

`s`	Total number of successes.
`n1`	Stage 1 sample size.
`r1`	Stage 1 critical value (trial is stopped at stage 1 if the number of successes is at most `r1`).
`n`	Total sample size.
`p0`	Response rate under the null hypothesis.

Details

Median unbiased estimator is the value of response rate such that the p-value is 0.5 (*Koyama and Chen, 2008*). The solution is found using numerical search, with a precision of 0.000001.

Value

Estimate of the response rate.

Author(s)

Arsenio Nhacolo

References

Koyama, T. and Chen, H. Proper inference from Simon's two-stage designs. *Stat Med*, 2008, 27, 3145-3154.

See Also

`pvaluek`, `pquantile`, `pm`, `pg`, `pu` and `pp`.

Examples

```
pk(21, 19, 4, 54, 0.2)
```

pm *Sample proportion*

Description

Calculates the sample proportion.

Usage

```
pm(s, n)
```

Arguments

s	Total number of successes.
n	Total sample size.

Details

For fixed designs the sample proportion is an unbiased (maximum likelihood) estimator of the response rate, but in group sequential designs (e.g., Simon's) it is biased.

Value

Estimate of the response rate.

Author(s)

Arsenio Nhacolo

See Also

pg, pu, pp and pk.

Examples

```
pm(21, 54)
```

pp *UMVCUE*

Description

Calculates the uniformly minimum variance conditionally unbiased estimator (UMVCUE) of the true response probability.

Usage

```
pp(s, n1, r1, n)
```

Arguments

`s`	Total number of successes.
`n1`	Stage 1 sample size.
`r1`	Stage 1 critical value (trial is stopped at stage 1 if the number of successes is at most `r1`).
`n`	Total sample size.

Details

The UMVCUE (*Pepe et al., 2009*) is conditional on on proceeding to the second stage. he sample proportion is used when the trial stopped at first stage.

Value

Estimate of the response rate.

Author(s)

Arsenio Nhacolo

References

Pepe, M. S.; Feng, Z.; Longton, G. and Koopmeiners, J. Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility. *Stat Med*, 2009, 28, 762-779.

See Also

`pm`, `pg`, `pu` and `pk`.

Examples

```
pp(21, 19, 4, 54)
```

`pquantile` *Value of response rate to attain a given p-value*

Description

Finds, for Simon-like designs, the value of response probability that would yield a given p-value.

Usage

```
pquantile(s, n1, r1, n, p0, pvalue)
```

Arguments

`s`	Total number of successes.
`n1`	Stage 1 sample size.
`r1`	Stage 1 critical value (trial is stopped at stage 1 if the number of successes is at most `r1`).
`n`	Total sample size.
`p0`	Response rate under the null hypothesis.
`pvalue`	The desired p-value.

Details

The solution is found using numerical search, with a precision of 0.000001. The p-value is as defined by *Koyama and Chen (2008)*.

Value

Response probability.

Author(s)

Arsenio Nhacolo

References

Koyama, T. and Chen, H. Proper inference from Simon's two-stage designs. *Stat Med*, 2008, 27, 3145-3154.

See Also

`pvaluek` and `pk`.

Examples

```
pquantile(21, 19, 4, 54, 0.2, 0.5)
```

pu *UMVUE*

Description

Calculates the uniformly minimum variance unbiased estimator (UMVUE) of the true response probability.

Usage

```
pu(s, n1, r1, n)
```

Arguments

s	Total number of successes.
n1	Stage 1 sample size.
r1	Stage 1 critical value (trial is stopped at stage 1 if the number of successes is at most r1).
n	Total sample size.

Details

The UMVUE is based on approach by *Grishick et al. (1946).* It was first considered by *Chang et al. (1989)* and further studied by *Jung et al. (2004).*

Value

Estimate of the response rate.

Author(s)

Arsenio Nhacolo

References

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

See Also

pm, pg, pp and pk.

Examples

```
pu(21, 19, 4, 54)
```

pvaluek *P-value*

Description

Calculates p-value for Simon-like designs.

Usage

```
pvaluek(s, n1, r1, n, p0)
```

Arguments

s	Total number of successes.
n1	Stage 1 sample size.
r1	Stage 1 critical value (trial is stopped at stage 1 if the number of successes is at most r1).
n	Total sample size.
p0	Response rate under the null hypothesis.

Details

It is based on the definition of p-value by *Koyama and Chen (2008)*.

Value

P-value.

Author(s)

Arsenio Nhacolo

References

Koyama, T. and Chen, H. Proper inference from Simon's two-stage designs. *Stat Med*, 2008, 27, 3145-3154.

See Also

`pquantile` and `pk`.

Examples

```
pvaluek(21, 19, 4, 54, 0.2)
```

`Pwr` *Power for single-stage parallel-group RCT.*

Description

`Pwr` calculates the power in an equal-size group two-arm randomized clinical trial with a binary response.

Usage

```
Pwr(pc, pt, Nc, alp = 0.05)
```

Arguments

`pc`	Response probability in control group.
`pt`	Response probability in treatment group.
`Nc`	Sample size per group.
`alp`	Significance level (default: 0.05).

Value

Sample size for one group.

Author(s)

Arsenio Nhacolo

References

Ahn, C., Heo, M. and Zhang, S. *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. CRC Press, 2014.

See Also

`Nct`.

Examples

```
Pwr(0.2,0.3,389,0.05)
```

`sbias` *Bias of the sample proportion*

Description

Calculates bias due to using sample proportion as estimator of the true response rate.

Usage

```
sbias(n1, r1, n, p)
```

Arguments

`n1`	Stage 1 sample size.
`r1`	Stage 1 critical value (trial is stopped at stage 1 if the number of successes is at most `r1`).
`n`	Total sample size.
`p`	True success probability.

Details

For fixed designs the sample proportion is an unbiased (maximum likelihood) estimator of the response rate, but in group sequential designs (e.g., Simon's) it is biased.

Value

Bias.

Author(s)

Arsenio Nhacolo

References

Porcher, R. and Desseaux, K. What inference for two-stage phase II trials? *BMC Med Res Methodol*, 2012, 12, 117.

See Also

`sfms` and `pg`.

Examples

```
sbias(19, 4, 54, 0.4)
```

sfms — *Probability mass function of* (M, S)

Description

Probability mass function of M (stage) and S (number of successes).

Usage

```
sfms(s, n1, r1, n, p, m = NA)
```

Arguments

s	Total number of successes.
n1	Stage 1 sample size.
r1	Stage 1 critical value (trial is stopped at stage 1 if the number of successes is at most r1).
n	Total sample size.
p	True success probability.
m	Stage number (1 or 2). It is automatically determined based on s and r1, therefore it shouldn't be provided, unless there are reasons to do so.

Details

Probability mass function of the statistic (M, S) for Simon-like designs (allowing early stopping for futility only).

Value

Density.

Author(s)

Arsenio Nhacolo

References

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

See Also

sbias and pg.

Examples

```
sfms(21, 19, 4, 54, 0.4)
```

SimulateEKOAD *Simulate single-arm binary endpoint two-stage adaptive designs.*

Description

SimulateEKOAD Simulate trials following designs similar to that of Englert and Kieser(2013)'s.

Usage

```
SimulateEKOAD(replicates, dsgn, newpi1 = NULL, seed = NULL,
  deleteOld = TRUE)
```

Arguments

replicates	Number of trials to be simulated.
dsgn	Dataframe containing one of the designs in EKOADwn.
newpi1	New response rate under the alternative hypothesis used to simulate trials. If NULL (default), the one from the design is used.
seed	The seed for random number generator. If NULL (default), no seed is set and , hence, results are not reproducible.
deleteOld	If TRUE (default), the simulation sub-directory is cleared before simulations start.

Details

The original designs (like the ones in EKOptAdaptDesigns) must be pre-processed using the function dsgnPrep to get extra information like the designs in EKOADwn.

Value

Simulated trials are saved in the sub-directory ./SimulatedTrials.

Author(s)

Arsenio Nhacolo

References

Englert, S. and Kieser, M. Optimal adaptive two-stage designs for phase II cancer clinical trials. *Biometrical Journal*, 2013.

See Also

EKOptAdaptDesigns, EKOADwn.

`SimulateSimonDsgn` *Simon's designs data simulation*

Description

`SimulateSimonDsgn` simulates data from Simon's optimal and minimax designs.

Usage

```
SimulateSimonDsgn(replicates, designParam, newp1 = NA, seed = NA,
  deleteOld = TRUE)
```

Arguments

`replicates` Number of trials to be generated.

`designParam` A dataframe containing Simon's optimal and minimax designs, as returned by the function `CalculateSimonDsgn`.

`newp1` If `NA` (default) data are generated assuming the same response probability under alternative hypothesis, `p1`, used to get the designs (see `CalculateSimonDsgn`). One may provide different values of `newp1` if there is interest in studying the effect of departure from the design's assumed `p1`.

`seed` Initial value (any integer) of random-number seed. It is useful for creating simulations that can be reproduced. The default is `NA`, meaning no reproducibility.

`deleteOld` If TRUE (default) the sub-directories `/Optimal/SimulatedTrials` and `/Minimax/SimulatedTri` are deleted, if they exist, before simulation starts. The old data files are still replaced by the new ones even if `deleteOld` is set to `FALSE`, but some old files remain in cases where the previous `replicates` was greater that the current one.

Details

The simulated trials are stored in the sub-directories `/Optimal/SimulatedTrials` and `/Minimax/SimulatedTrials` for optimal and minimax designs, respectively, under the current working directory. The sub-directories are automatically created. Individual trial data are stored in a CSV file named `trial#`, where # is the replicate number.

Value

The function is not intended to return an R object, instead it creates files (in CSV format) containing simulated trials data. See *Details*. It also saves in the current working directory the `designParam` argument (*DesignParameters.csv*).

Author(s)

Arsenio Nhacolo

See Also

`CalculateSimonDsgn`, `SimulateSimonDsgnAdaptN` and `AnalyzeSimonDsgn`.

Examples

```
d <- CalculateSimonDsgn(0.2, 0.4, 0.05, 0.1)
SimulateSimonDsgn(100, d, seed = 1986)
```

SimulateSimonDsgnAdaptN

Simon's adaptive designs data simulation

Description

Simulates data from adaptive versions of Simon's optimal and minimax designs, proposed by *Englert and Kieser (2012)*. Adaptation consists in recalculating the second stage sample size n2 in order to achieve a desired conditional power given the number of successes at first stage.

Usage

```
SimulateSimonDsgnAdaptN(replicates, designParam, newp1 = NA,
  condPwr = NA, restAlphaMet = 0, seed = NA, deleteOld = TRUE)
```

Arguments

`replicates` Number of trials to be generated.

`designParam` A dataframe containing Simon's optimal and minimax designs, as returned by the function `CalculateSimonDsgn`.

`newp1` If `NA` (default) data are generated assuming the same response probability under alternative hypothesis, p1, used to get the designs (see `CalculateSimonDsgn`). One may provide different values of `newp1` if there is interest in studying the effect of departure from the design's assumed `p1`.

`condPwr` The desired conditional power. The default is `1-beta`.

`restAlphaMet` The method for spending the "rest alpha" (difference between nominal alpha level and actual alpha level for the given design).

- 0: "rest alpha" is not used (default);
- 1: "rest alpha" is spent proportionally;
- 2: "rest alpha" is spent equally;
- 3: "rest alpha" is spent only to the worst case scenario (minimal number of responses at the interim analysis so that the study can proceed to the second stage).

`seed` Initial value (any integer) of random-number seed. It is useful for creating simulations that can be reproduced. The default is `NA`, meaning no reproducibility.

`deleteOld` If TRUE (default) the sub-directories `/OptimalAdapt/SimulatedTrials` and `/MinimaxAdapt/SimulatedTrials` are deleted, if they exist, before simulation starts. The old data files are still replaced by the new ones even if `deleteOld` is set to `FALSE`, but some old files remain in cases where the previous `replicates` was greater that the current one.

Details

The simulated trials are stored in the sub-directories `/OptimalAdapt/SimulatedTrials` and `/MinimaxAdapt/Simula` for optimal and minimax designs, respectively, under the current working directory. The sub-directories are automatically created. Individual trial data are stored in a CSV file named `trial#`, where # is the replicate number.

Value

The function is not intended to return an R object, instead it creates files (in CSV format) containing simulated trials data. See *Details*. It also saves in the current working directory the `designParam` argument (*DesignParametersAdapt.csv*).

Author(s)

Arsenio Nhacolo

References

Englert S., Kieser M. Adaptive designs for single-arm phase II trials in oncology. *Pharm Stat*, 2012, 11, 241-249.

See Also

`CalculateSimonDsgn`, `getN2`, `SimulateSimonDsgn` and `AnalyzeSimonDsgnAdaptN`.

Examples

```
d <- CalculateSimonDsgn(0.2, 0.4, 0.05, 0.1)
SimulateSimonDsgnAdaptN(100, d, seed = 1986)
```

Index

Appendix D

First paper

Check for updates

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
0(0) 1–14
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280218781411
journals.sagepub.com/home/smm
SAGE

Interval and point estimation in adaptive Phase II trials with binary endpoint

Arsénio Nhacolo and Werner Brannath

Abstract

Phase II clinical trials are concerned with making decision of whether a treatment is sufficiently efficacious to be worth further investigations in late large scale Phase III trials. In oncology Phase II trials, frequentist single-arm two-stage group-sequential designs with a binary endpoint are commonly used. To allow for more flexibility, adaptive versions of these designs have been proposed. In this paper, we propose point and interval estimation for adaptive designs in which the second stage sample size is a pre-specified function of first stage's number of responses. Our approach is based on sample space orderings, from which we derive *p*-values, and point and interval estimates. Simulation studies show that our proposed methods perform better, in terms of bias and root mean square error, than the fixed-sample maximum likelihood estimator.

Keywords

Adaptive designs, bias, clinical trials, estimation, oncology Phase II, RMSE, *p*-value

1 Introduction

Phase II trials are concerned with making decision of whether a treatment is sufficiently efficacious to justify its further investigations in late large scale Phase III trials. In oncology Phase II trials, frequentist single-arm two-stage group-sequential designs with binary endpoints are commonly used. Based on ethical desirability to expose less patients to an inefficient treatment and to speed-up the development process, these designs allow early termination of the trial for futility and/or efficiency (e.g., designs by Schultz et al.[1] and Simon[2]). In such designs, the sample sizes and decision rules for each stage are predefined. To allow flexibility, adaptive versions of these designs have been proposed.[3–10] Adaptive designs allow for modification of the trial at interim analysis using the trial's accumulating data and/or external information without jeopardising trial's integrity and validity.

Although the main goal of oncology Phase II trials is hypothesis testing, estimation of the efficacy parameter after such trials remains important, especially in cases where the treatment was deemed successful since it will be needed for planning Phase III trials. In group-sequential designs (GSD), due to the possibility of early stopping for either futility or efficacy, the fixed-sample maximum likelihood estimator (MLE) of the treatment effect (response probability) is no longer unbiased. This issue has been acknowledged by many authors, and alternative estimation methods have been proposed in the literature.[11–19] However, most of the estimation methods for GSD are not applicable to adaptive designs. Unfortunately, the literature on estimation in adaptive GSD is mainly on Phase III clinical trials.[20–35] To the best of our knowledge, estimation in oncology Phase II adaptive single-arm GSD with a binary endpoint has only been discussed recently by Kunzmann and Kieser,[36] who proposed a point estimator that can be interpreted as a constrained posterior mean estimate based on the non-informative Jeffreys prior. Their method is computationally intensive, and they have implemented it in the Julia[37] programming language using the JuMP[38] package and the Julia interface[38] to the commercial solver Gurobi.[39]

In this paper, as an alternative to the Bayesian procedure by Kunzmann and Kieser,[36] we propose a frequentist interval and point estimation procedure for two-stage single-arm adaptive designs with pre-specified adaptation rule. We consider designs in which the second stage's sample size is a pre-specified function of first stage's number

Competence Centre for Clinical Trials, University of Bremen, Bremen, Germany

Corresponding author:
Arsénio Nhacolo, Competence Centre for Clinical Trials, University of Bremen, Linzer Straße 4, Raum 41010, Bremen 28359, Germany.
Email: anhacolo@uni-bremen.de

of responses. However, some of the approaches that we propose can be extended to flexible designs. The procedure uses the concept of stage-wise ordering, and it is less computationally intensive and can readily be implemented in the statistical programming language R.[40] We first propose and discuss different approaches for defining sample space orderings, from which we derive p-values and then interval and point estimates. We also present the results from a simulation study to evaluate the performance of the proposed methods. The paper is organised as follows. We first give an overview of the adaptive designs for which we are developing the proposed methods. Afterwards we give the methodological details of our proposals, an illustrative example, then the results of the simulation study and we end with conclusions and a discussion.

2 Adaptive phase II oncology designs

We build our methodology for two-stage adaptive phase II oncology designs with binary endpoint and a pre-specified adaptation rule, like those proposed by Englert and Kieser[8] and Shan et al.[9] These designs extend the classical binary endpoint oncology Phase II GSD by allowing the sample size of second stage to depend on the number of responses observed in the first stage. Like their classical GSD counterparts they test, at type I error rate α and type II error rate β, the null (H_0) versus the alternative (H_1) hypothesis about the response rate (π)

$$H_0 : \pi \leq \pi_0 \text{ vs } H_1 : \pi \geq \pi_1$$

where π_0 is the maximum response rate considered to be uninteresting and π_1 is the minimum desirable response rate, with $\pi_1 > \pi_0$.

The designs we consider in this paper are defined by the first stage sample size, n_1, futility and efficacy boundaries, l_1 and u_1 ($u_1 > l_1$), which are fixed, and the second stage sample size, $n_2(x_1)$, which depends on the number of responses observed in the first stage, x_1. We further have the conditional error function, $D(x_1)$, and the corresponding decision boundary, $l(x_1)$, which are also functions of x_1. The final (second) stage efficacy boundary $u(x_1)$ is set to $u(x_1) = l(x_1) + 1$, with $l(x_1)$ being the futility boundary. $D(x_1)$ defines for each possible number of responses in the first stage, $x_1 \in \{0, \ldots, n_1\}$, the conditional type I error rate to be used in the second stage.[6] At the interim analysis, the trial is stopped with failure to reject H_0 if $x_1 \leq l_1$ or with rejection of H_0 if $x_1 \geq u_1$. Otherwise the trial proceeds to the second (final) stage, after which H_0 is rejected if $p_2 < D(x_1)$ or, equivalently, $x > l(x_1)$, where p_2 is the second stage p-value and x is the total number of responses (i.e. x is the sum of x_1 and the number of responses observed in the second stage, x_2). Note that $x > l(x_1)$ is equivalent to $x \geq u(x_1)$. An example of such designs is given in Table 1.

Note that the discrete conditional error function $D(x_1)$ in Table 1 is non-decreasing in x_1, and takes values within $[0, 1]$. We assume these two properties throughout the paper. It is clear that in these designs the first stage decision boundaries are $l_1 = \max\{x_1 | D(x_1) = 0\} = \min\{x_1 | D(x_1) > 0\} - 1$ and $u_1 = \min\{x_1 | D(x_1) = 1\} = \max\{x_1 | D(x_1) < 1\} + 1$, and the first and second stage p-values are, respectively, $p_1 = 1 - B(x_1 - 1, n_1, \pi_0)$ and $p_2 = 1 - B(x_2 - 1, n_2(x_1), \pi_0)$, where $B(x, n, \pi)$ is the binomial cumulative distribution function (c.d.f) with x successes, n trials and success probability π.

Table 1. Englert and Kieser's[8] optimal adaptive design for $(\pi_0, \pi_1, \alpha, \beta) = (0.2, 0.4, 0.05, 0.1)$.

$n_1 = 20, n_{2,max} = 39$

x_1	$n_2(x_1)$	$D(x_1)$	$l(x_1)$
≤4	0	0	0
5	16	0.082	10
6	30	0.129	14
7	33	0.200	15
8	39	0.241	17
9	39	0.376	17
≥10	0	1	0

3 Classical sample space orderings

The construction of confidence intervals and p-values entails determining the probability of obtaining an outcome that is at least as extreme as the observed one and, for this, a sample space ordering is needed. For the case of one outcome variable, in fixed-sample designs, the sample space ordering is simply the ordering of the real numbers. However, in GSD the ordering is not clearly defined because, apart from the test statistic, the number of stages also plays a role when ordering the outcomes. Different sample space orderings for GSD have been suggested in the literature. The stage-wise ordering, first proposed by Armitage[41] and later discussed by several other authors,[42–46] is a widely used sample space ordering in GSD. For classical GSD counterparts of the adaptive designs above, i.e. designs in which n_2 and l are also fixed, the stage-wise ordering can be defined as it follows. Let m be the stopping stage and x the total number of responses. A trial outcome (m', x') is at least as extreme (against H_0) as the observed trial outcome $(m,\ x)$, written as $(m', x') \succcurlyeq (m, x)$, if one of the following conditions is met

$$\begin{aligned} &(\mathrm{A}) \quad m' = m \text{ and } x' \geq x \\ &(\mathrm{B}) \quad m' = 1,\ m = 2 \text{ and } x' \geq u_1 \\ &(\mathrm{C}) \quad m' = 2,\ m = 1 \text{ and } x \leq l_1 \end{aligned}$$

Other suggested sample space orderings include the likelihood ratio ordering,[47–49] sample mean ordering,[50] and score test ordering.[49]

For the adaptive designs, the stage-wise ordering discussed here can be inconsistent with the design's decision rule when $m' = m = 2$. For example, for the design in Table 1, $x = 11$ with $x_1 = 5$ leads to rejection of H_0 while $x = 16$ with $x_1 = 8$ does not. This follows from the nature of the conditional error function.

4 Alternative sample space orderings

To overcome this inconsistency, we propose alternative sample space orderings that take into account the conditional error function and adaptation rule. When both outcomes are from trials that continued to the second stage (i.e. $m' = m = 2$), we compare them taking into account their respective rejection boundaries. We accomplish this by defining a function $\delta(x_1, x_2)$ that in some way incorporates the rejection boundary of the trial outcome. In all other cases, the proposed sample space orderings are similar to the stage-wise ordering discussed above. Then we have that $(m', x_1', x') \succcurlyeq (m, x_1, x)$ if one of the following conditions is met

$$\begin{aligned} &(\mathrm{A1}) \quad m' = m = 1 \text{ and } x' \geq x \\ &(\mathrm{A2}) \quad m' = m = 2 \text{ and } \delta(x_1', x_2') \geq \delta(x_1, x_2) \\ &(\mathrm{B}) \quad m' = 1,\ m = 2 \text{ and } x' \geq u_1 \\ &(\mathrm{C}) \quad m' = 2,\ m = 1 \text{ and } x \leq l_1 \end{aligned}$$

We propose three different methods to define $\delta(x_1, x_2)$. The first two quantify the deviation between x and $l(x_1)$. In the first method, we define $\delta(x_1, x_2)$ using directly x as

$$\delta(x_1, x_2) = x_1 + x_2 - l(x_1) = x - l(x_1) \tag{1}$$

and in the second we define $\delta(x_1, x_2)$ using the second stage p-value as

$$\delta(x_1, x_2) = \tilde{\delta}[x_1, p_2(x_2)] = D(x_1) - p_2(x_2) \tag{2}$$

In both methods, δ is defined such that it equals to a constant when the outcome is at the decision boundary (i.e. when $x_2 = l(x_1) - x_1$ and $p_2 = D(x_1)$). That is, $\delta[x_1, l(x_1) - x_1] = c_1$ and $\delta[x_1, D(x_1)] = c_2$. Here $c_1 = c_2 = 0$, meaning that the null hypothesis is rejected if $\delta(x_1, x_2) > 0$. The inequality $\delta(x_1', x_2') \geq \delta(x_1, x_2)$ in the case (A2) of our proposed sample space ordering can be stated as $x_2' \geq x - l(x_1) + l(x_1') - x_1'$ for the first method and $p_2' \leq p_2 - D(x_1) + D(x_1')$ for the second one.

The two ways of defining the function δ above are strictly linked to the design's decision rules, and therefore require trials to strictly follow the design. We will see later that Method 2 yields valid p-values even if the adaptation rule is not strictly adhered to. One way to allow for flexibility is to order the outcomes using combination functions from adaptive tests. Combination functions combine the first and the second stage p-values, with the assumption that the data from the two stages are from independent cohorts of patients. An extensive discussion on adaptive combination tests can be found in Wassmer and Brannath.[46] The idea is to define a combination function $C(p_1, p_2)$, setting $\alpha_0 = 1 - B(l_1 - 1, n_1, \pi_0)$, $\alpha_1 = 1 - B(u_1 - 1, n_1, \pi_0)$, and finding c such that type I error is controlled, i.e.

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 I_{[C(p_1,p_2) \le c]} d_{p_2} d_{p_1} = \alpha$$

where $I_{[S]}$ equals to 1 if S is true and 0 otherwise.

Given the combination test, the most natural ordering on the second stage is according to $C(p_1, p_2)$, i.e. if we have performed a second stage we consider a second trial outcome with stage-wise p-values (p_1', p_2') as more extreme than our observed outcome if $C(p_1', p_2') < C(p_1, p_2)$. Even though the Phase II designs we are dealing with might not be based on a combination function C, we can build an ordering based on C that is consistent with the rejection region given by the function D (or equivalently given by the function l). To this end, we define the C's corresponding conditional error function

$$A(p_1) = \max\{y \in [0, 1] : C(p_1, p_2) \le c\}$$

and then calculate the backwards image p_{1b} such that $A(p_{1b}) = D(x_1)$, where $D(x_1)$ is the conditional error of the original design.

A natural and common choice for $C(p_1, p_2)$ is the weighted inverse normal combination function,[51] which can be represented as[46]

$$C(p_1, p_2) = 1 - \Phi\left[w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)\right]$$

where Φ is standard normal c.d.f., and w_1 and w_2 are predefined weights chosen such that $w_1^2 + w_2^2 = 1$. Here we propose to use weights that give more emphasis to the stage with higher sample size, i.e.

$$w_1 = \sqrt{\frac{n_1}{n_1 + n_2(x_1)}} \quad \text{and } w_2 = \sqrt{\frac{n_2(x_1)}{n_1 + n_2(x_1)}}$$

The conditional error function of the inverse normal combination function is

$$A(p_1) = 1 - \Phi\left[\frac{\Phi^{-1}(1 - c) - w_2 \Phi^{-1}(1 - p_2)}{w_1}\right]$$

Solving $A(p_{1b}) = D(x_1)$ for p_{1b} we get

$$p_{1b}(x_1) = 1 - \Phi\left\{\frac{\Phi^{-1}(1 - c) - w_2 \Phi^{-1}[1 - D(x_1)]}{w_1}\right\}$$

We finally define δ as

$$\delta(x_1, x_2) = \bar{\delta}[p_{1b}(x_1), p_2(x_2)] = 1 - C(p_{1b}, p_2) \tag{3}$$

With δ defined in this way, the condition $\delta(x_1', x_2') \ge \delta(x_1, x_2)$ in the case (A2) of the proposed sample space ordering becomes $C(p_1' b, p_2') \le C(p_{1b}, p_2)$, meaning that the outcome with lower $C(p_{1b}, p_2)$ is considered to be more extreme.

Another possible sample space ordering could be to simply order the trial outcomes by the proportion of responses, i.e. $(x_1 + x_2)/[n_1 + n_2(x_1)]$. However, this ordering would not always be consistent with the design's decision rule because in some designs the ratio of the final decision boundary and the total sample size, i.e. $l(x_1)/[n_1 + n_2(x_1)]$, might not be constant. For instance, in the Englert and Kieser's[8] design for $(\pi_0, \pi_1, \alpha, \beta) = (0.5, 0.7, 0.05, 0.2)$, this ratio is 0.6 if $x_1 = 14$ and 0.8 if $x_1 = 15$.

5 Overall *p*-value

We use the sample space ordering proposed in the previous section to derive an overall p-value, denoted by Q, meant to be calculated when the trial has been terminated. Q is defined as the probability of observing under H_0 an outcome (m', x'_1, x') that is similar or more extreme than the outcome (m, x_1, x) actually observed in the trial. If the observed outcome is from a trial that stopped at the first stage, outcomes with $x'_1 \geq x_1$ are more extreme, irrespective of their stopping stage, implying the overall p-value

$$Q = \Pr_{\pi_0}(X_1 \geq x_1)$$

If the observed outcome is from a trial that continued to the second stage, more extreme are outcomes from trials that stopped at the first stage with $x'_1 \geq u_1$ or continued to the second stage with $\delta(x'_1, x'_2) \geq \delta(x_1, x_2)$, then

$$Q = \Pr_{\pi_0}(X_1 \geq u_1) + \sum_{x'_1 = l_1+1}^{u_1-1} \Pr_{\pi_0}(X_1 = x'_1) \Pr_{\pi_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | X_1 = x'_1\right]$$

Since X_1 and X_2 follow binomial distribution, we can write the overall p-value as

$$Q = \begin{cases} 1 - B(x_1 - 1, n_1, \pi_0) & \text{if } m = 1 \\ 1 - B(u_1 - 1, n_1, \pi_0) & \\ + \sum\limits_{x'_1 = l_1+1}^{u_1-1} b(x'_1, n_1, \pi_0) \Pr_{\pi_0}\left(\Delta \geq \delta | x'_1\right) & \text{if } m = 2 \end{cases}$$

where $b(x, n, \pi)$ is the binomial probability mass function with x successes, n trials and success probability π, $\Delta = \delta(X_1, X_2)$ and $\delta = \delta(x_1, x_2)$.

We discuss in the following lines approaches to calculate the probability of $\delta(X_1, X_2) \geq \delta(x_1, x_2)$ under H_0, i.e. $\Pr_{\pi_0}[\delta(X_1, X_2) \geq \delta(x_1, x_2)]$. For the $\delta(x_1, x_2)$ defined in equation (1), since we are working directly with the number of events (responses), this probability can easily be calculated as (we call this *Method 1*)

$$\begin{aligned} \Pr_{\pi_0}\left(\Delta \geq \delta | x'_1\right) &= \Pr_{\pi_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | x'_1\right] \\ &= \Pr_{\pi_0}\left[X - l(X_1) \geq x - l(x_1) | x'_1\right] \\ &= \Pr_{\pi_0}\left[X_1 + X_2 - l(X_1) \geq x - l(x_1) | x'_1\right] \\ &= \Pr_{\pi_0}\left[X_2 \geq x - l(x_1) + l(X_1) - X_1 | x'_1\right] \\ &= 1 - B\left[x - l(x_1) + l(x'_1) - x'_1 - 1, n_2(x'_1), \pi_0\right] \end{aligned}$$

For the other two methods, we use approximations. We make use of the fact that a p-value P is in general stochastically not smaller than a standard uniform variate, i.e.

$$\Pr_{\pi_0}(P \leq \gamma) \leq \gamma, \quad \gamma \in [0, 1]$$

Assuming that the first stage design is pre-fixed and is strictly followed, and that the first and the second stage data are from independent cohorts of patients, using the second stage p-value p_2 as the test statistic guarantees the *conditional invariance principle*.[46] Conditional on the first stage data and second stage design, the distribution of p_2 under H_0 is not smaller than the uniform distribution. This implies that the type I error rate is controlled irrespective of the adaptation rule.

When using the $\delta(x_1, x_2)$ in (2) (*Method 2*), we approximate $\Pr_{\pi_0}(\Delta \geq \delta)$ as

$$\begin{aligned} Pr_{\pi_0}\left(\Delta \geq \delta | x'_1\right) &= \Pr_{\pi_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | x'_1\right] \\ &= \Pr_{\pi_0}\left[D(X_1) - P_2 \geq D(x_1) - p_2 | x'_1\right] \\ &= \Pr_{\pi_0}\left[P_2 \leq p_2 - D(x_1) + D(X_1) | x'_1\right] \\ &\approx \left\langle p_2 - D(x_1) + D(x'_1)\right\rangle_{[0,1]} \end{aligned}$$

where

$$\langle \omega \rangle_{[0,1]} = \begin{cases} 0 & \text{if } \omega < 0 \\ \omega & \text{if } 0 \leq \omega \leq 1 \\ 1 & \text{if } \omega > 1 \end{cases}$$

Another way of calculating $Pr_{\pi_0}(\Delta \geq \delta | x_1')$ in Method 2, denoted *Method 2v2*, is to use the fact that $p_2 = 1 - B[x_2 - 1, n_2(x_1), \pi_0]$ and the binomial quantile function, denoted by B_q, as it follows

$$\begin{aligned} \Pr_{\pi_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | x_1'\right] &= \Pr_{\pi_0}\left[P_2 \leq p_2 - D(x_1) + D(X_1) | x_1'\right] \\ &= \Pr_{\pi_0}\left\{1 - B[X_2 - 1, n_2(X_1), \pi_0] \leq p_2 - D(x_1) + D(X_1)\right\} \\ &= \Pr_{\pi_0}\left\{B[X_2 - 1, n_2(X_1), \pi_0] \geq 1 - p_2 + D(x_1) - D(X_1)\right\} \\ &= \Pr_{\pi_0}\left\{X_2 - 1 \geq B_q[1 - p_2 + D(x_1) - D(X_1), n_2(X_1), \pi_0]\right\} \\ &= \Pr_{\pi_0}\left\{X_2 \geq 1 + B_q[1 - p_2 + D(x_1) - D(X_1), n_2(X_1), \pi_0]\right\} \\ &= 1 - B\left\{B_q\left[1 - p_2 + D(x_1) - D(x_1'), n_2(x_1'), \pi_0\right], n_2(x_1'), \pi_0\right\} \end{aligned}$$

Finally, using the $\delta(x_1, x_2)$ in (3), *Method 3*, we have that

$$\begin{aligned} \Pr_{\pi_0}\left(\Delta \geq \delta | x_1'\right) &= \Pr_{\pi_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | x_1'\right] \\ &= \Pr_{\pi_0}\left[C(P_1 b, P_2) \leq C(p_{1b}, p_2) | x_1'\right] \\ &= \Pr_{\pi_0}\left[P_2 \leq 1 - \Phi(z_b) | x_1'\right] \\ &\approx 1 - \Phi(z_b) \end{aligned}$$

where

$$z_b = \frac{w_1 \Phi^{-1}(1 - p_{1b}) + w_2 \Phi^{-1}(1 - p_2) - w_1' \Phi^{-1}(1 - p_1' b)}{w_2'}$$

See more details in Appendix 1.

6 Point and interval estimation

We follow the approach discussed in Chapter 8 of Wassmer and Brannath.[46] Exploiting the duality between confidence intervals (CI) and hypothesis tests, we construct the CI by considering all the null hypotheses

$$H_0^{\tilde{\pi}_0} : \pi \leq \tilde{\pi}_0, \text{ with } 0 \leq \tilde{\pi}_0 \leq 1$$

The confidence set is a collection of $\tilde{\pi}_0$ for which H_0 is not rejected. Assuming that the overall p-value Q is monotone increasing in $\tilde{\pi}_0$ for all outcomes (m, x_1, x), the region $\left\{\tilde{\pi}_0 : Q(\tilde{\pi}_0) = \Pr_{\tilde{\pi}_0}[(M, X_1, X) \succcurlyeq (m, x_1, x)] > \alpha\right\}$ is a one-sided $(1 - \alpha)100\%$ CI defined as $\left]\pi_L^\alpha; 1\right]$, where the lower bound π_L^α is the solution, in $\tilde{\pi}_0$, of the equation $Q(\tilde{\pi}_0) = \alpha$.

As the point estimate we take the lower bound of the 50% one-sided CI, i.e. $\hat{\pi} = \pi_L^{0.5}$, which is an approximate median unbiased estimator. Similar estimators have been proposed for classical oncology two-stage GSDs by Koyama and Chen[14] and Jovic and Whitehead,[18] which are applicable only if n_2 is a constant for all x_1.

There are four different methods (Method 1, Method 2, Method 2v2 and Method 3) for calculating the overall p-value Q (see the previous section) on which the estimation approach is based, resulting, therefore, in four different estimates. We name these estimates after their respective p-value calculation methods, to make a distinction amongst them throughout the rest of this paper. Table 2 lists the methods together with the function δ used in sample space ordering, the way $Pr_{\pi_0}(\Delta \geq \delta | x_1')$ (part of Q that differ across the methods) is estimated, and the implications for estimation.

Table 2. Methods for calculating the overall p-value Q.

Method	δ	$Pr_{\pi_0}(\Delta \geq \delta \vert x\prime_1)$	Implications for estimation
1	$x - l(x_1)$	$1 - B[x - l(x_1) + l(x\prime_1) - x\prime_1 - 1, n_2(x\prime_1), \pi_0]$	Exact
2	$D(x_1) - p_2(x_2)$	$\langle p_2 - D(x_1) + D(x\prime_1)\rangle_{[0,1]}$	Approximate
2v2	$D(x_1) - p_2(x_2)$	$1 - B\{B_q[1 - p_2 + D(x_1) - D(x\prime_1), n_2(x\prime_1), \pi_0], n_2(x\prime_1), \pi_0\}$	Exact
3	$1 - C(p_1 b, p_2)$	$1 - \Phi(z_b)$	Approximate

Note: δ denotes the function used for sample space ordering and $P_{r_{\pi_0}}(\Delta \geq \delta | x_1')$ is the part of the Q formula that vary across the methods. In the "Implications for estimation," *Exact* means the resulting point and interval estimates are obtained using exact probability calculations, while *Approximate* means approximations are used, i.e. the estimation is based on continuous uniform distribution of second stage p-value.

As mentioned before, in order to this estimation technique to work, it is necessary that the overall p-value Q as function of response probability π be monotone increasing for $\pi \in [0, 1]$. We checked the monotonicity of $Q(\pi)$ numerically for all 34 designs listed in Englert and Kieser,[8] for all possible outcomes and π ranging from 0 to 1 by increments of 0.01. We found that Methods 2 and 3 are monotone in all designs. Method 1 is monotone, except for four designs when $\pi \geq 0.8$. In the case of non-monotonicity, a conservative solution may be found using the cumulative maximum of $Q(\pi)$, i.e. $Q_{cm}(\pi) = \max\{Q(\pi') : \pi' \leq \pi\}$.

We give more details on how the estimation is done for each of the three methods in Appendix 1.

7 Illustrative example

Suppose that a Phase II trial testing the activity of an anti-cancer agent was conducted using the design given in Table 1. Suppose further that at the interim analysis it was found that 8 out of 20 patients responded to the treatment, leading to the decision to proceed with the trial to the second (final) stage. Therefore, 39 additional patients were recruited and treated, of which 18 were responsive. With a total of 26 responders out of 59 patients at the end of trial, the decision according to the design's decision rule was to reject the null. Calculating the overall p-value using the proposed approach, methods 1, 2, 2v2 and 3, we get the values 0.00261, 0.00360, 0.00315 and 0.00261, respectively. As it can be seen, all the p-values are less than the significance level $\alpha = 0.05$, meaning that with the proposed methods we also reject the null hypothesis. With the null hypothesis rejected, one would be interested in getting the response rate estimate to possibly use it for planing further trials. If we would ignore the adaptive nature of the design and employ the fixed-sample maximum likelihood estimator, the point estimate would be 0.44068. The point estimates obtained using the proposed approach are 0.42264, 0.41367, 0.41337 and 0.41411, respectively, for the methods 1, 2, 2v2 and 3. The corresponding 95% one-sided CIs are]0.29561; 1],]0.29105; 1],]0.27918; 1] and]0.29379; 1].

8 Numerical study

We did an extensive numerical study to evaluate various aspects of the proposed methods. We computed, for all designs in Englert and Kieser,[8] the overall p-value to see how it behaves for the three methods, as well as to check whether it is consistent with the original decision rule. The computation was for all possible outcomes, with values of π_0 varying from 0 to 1 by 0.01 for each. We then did a simulation study to assess the performance of the proposed estimator using the three methods. The performance of the point estimate was quantified in terms of bias and root mean square error (RMSE), and the performance of the interval estimate was quantified by the coverage probability and mean of the lower bound. We calculated, in addition, the type I error and power using the original decision rule and using the overall p-value from the proposed methods.

Bias was defined as $\frac{1}{T}\sum_{t=1}^{T}(\hat{\pi}_t - \pi)$ and RMSE as the square root of $\frac{1}{T}\sum_{t=1}^{T}(\hat{\pi}_t - \pi)^2$, where T is the total number of simulated trials, $\hat{\pi}$ the estimated response probability and π the response probability under which trials were simulated. The coverage probability was computed as the proportion of trials in which the $(1 - \alpha)100\%$ CI contained the true response rate π, i.e., proportion of trials in which the lower bound of the CI is less than π. The type I error was calculated as the proportion of trials simulated under $\pi = \pi_0$ in which H_0 was rejected, and the power calculated similarly but for trials simulated under $\pi = \pi_1$.

For comparison purposes, we included the maximum likelihood estimator (MLE) for fixed-sample designs, which we believe is more likely to be employed when analysing data from adaptive designs for which no specific

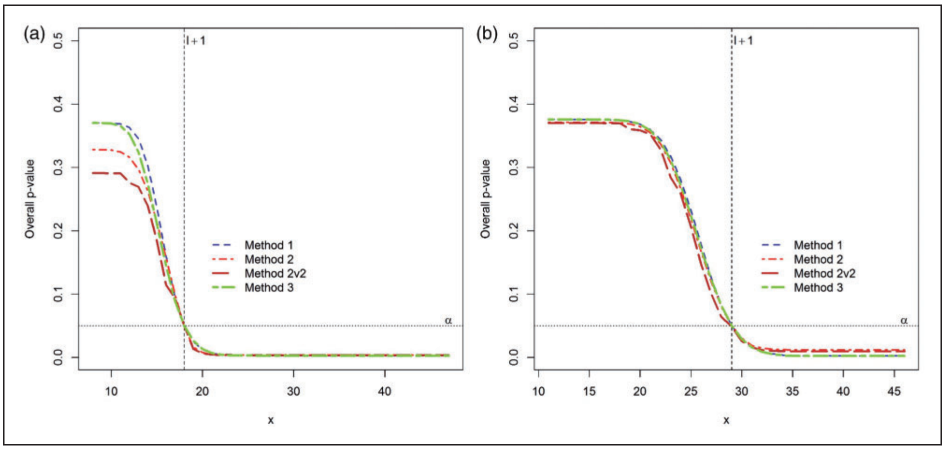


Figure 1. Plot of overall p-value (Q) as function of the total number of responses (x). (a) is of the design $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$ with $x_1 = 5$, and (b) of $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$ with $x_1 = 10$. Both are cases where the trial continues to second stage. The vertical line represents the minimum total number of responses necessary to reject H_0 using design's decision rule.

estimation methods are available. When applying this estimator, we ignored the adaptive nature of the design and we did estimation as if the data were from a fixed-sample (single-stage) design. Therefore, the MLE is the sample proportion of the pooled data. We didn't include the estimators proposed for classical GSDs mentioned above. Their formulae are based on the fact that the second stage sample size is pre-defined and constant, therefore, they would not be applicable here without modification. We used two versions of MLE, one that uses all trial data, $\hat{\pi}_p = [x_1 + x_2]/[n_1 + n_2(x_1)]$, and the other that uses the first stage data only, $\hat{\pi}_{p1} = x_1/n_1$. The reason for including $\hat{\pi}_{p1}$ is that since it is unbiased, it will serve as benchmark for comparison with respect to RMSE, i.e. a new estimator would not be desirable if it would be outperformed by $\hat{\pi}_{p1}$ in terms of RMSE. We denote the estimated response probability by $\hat{\pi}_{m1}$ for Method 1, $\hat{\pi}_{m2}$ for Method 2, $\hat{\pi}_{m2v2}$ for Method 2v2, and $\hat{\pi}_{m3}$ for Method 3. The simulation were done for two designs of Englert and Kieser,[8] one with a moderate π_1, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, we call this design 1, and the other (design 2) with relatively high π_1, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. For both designs we varied, in the simulated trials, the true response probability π from 0 to 1 by increments of 0.01. For each scenario, 50,000 simulations were run.

We implemented all the methodology described above in the statistical programming language R.[40] The code and datasets containing all designs listed in Englert and Kieser[8] are available from the authors upon request.

9 Results

Figure 1 shows plots of overall p-value for all possible values of x when $x_1 = 8$ and $x_1 = 11$ in designs 1 and 2, respectively. It can be seen that, as expected, all methods are consistent with design's decision rule. Results from simulation study in Table 3 reveal that type I error rate and power of Methods 1 and 2v2 are equal to those of design's original decision rule, which are in turn very close to the nominal levels. Methods 2 and 3 are conservative, as their type I error rate is lower compared to other methods.

Simulation results on bias and RMSE of the estimators for values of π ranging from 0 to 1 are shown in Figures 2 and 3, respectively. Table 4 shows the results of simulations under H_1 ($\pi = \pi_1$), and, in addition to bias and RMSE, it shows the mean and the first, second and third quartiles of the estimates, and the coverage probability and the mean lower bound of the one-sided $(1 - \alpha)100\%$ CI. The same results for simulations under $\pi = \pi_1 + 0.1$ are shown in Table 5. The behaviour of the estimators change depending on whether the true response rate (π) is close to or far from the hypothesised one (π_1). Taking a closer look at Figure 2 we can see that, except the first stage sample proportion ($\hat{\pi}_{p1}$) which is unbiased as expected, all estimators are negatively mean biased for values of π around π_0. When the true response rate is close to π_1, the estimators of the proposed methods ($\hat{\pi}_{m1}$, $\hat{\pi}_{m2}$, $\hat{\pi}_{m2v2}$ and $\hat{\pi}_{m3}$) are almost unbiased, while the fixed sample MLE ($\hat{\pi}_p$) shows positive bias. As π approaches 1, the proposed estimators become more and more negatively biased, while the bias of $\hat{\pi}_p$

Table 3. Type I error and power based on design's original decision rule (Orig.) and on the three proposed methods (Met.), from 50,000 simulation runs.

	Decision rule				
	Orig.	Met. 1	Met. 2	Met.2v2	Met. 3
Type I	0.0503	0.0503	0.0430	0.0503	0.0430
Power	0.9002	0.9002	0.8877	0.9002	0.8877
Type I	0.0492	0.0492	0.0421	0.0492	0.0428
Power	0.8990	0.8990	0.8884	0.8990	0.8892

Note: The two first rows are for design 1, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and the last two for the design 2, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$.

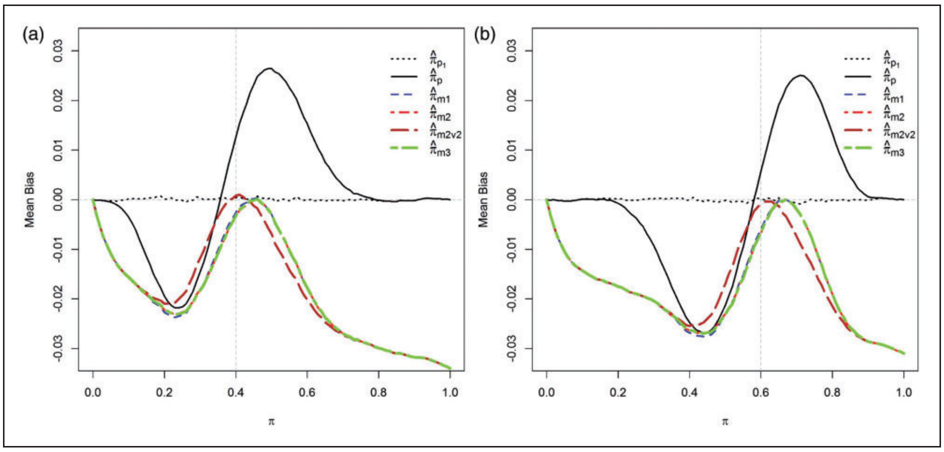


Figure 2. Mean bias of estimators. Design 1 is defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and 2 by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. For each value of π, 50,000 trials were simulated. The vertical line represents $\pi = \pi_1$.

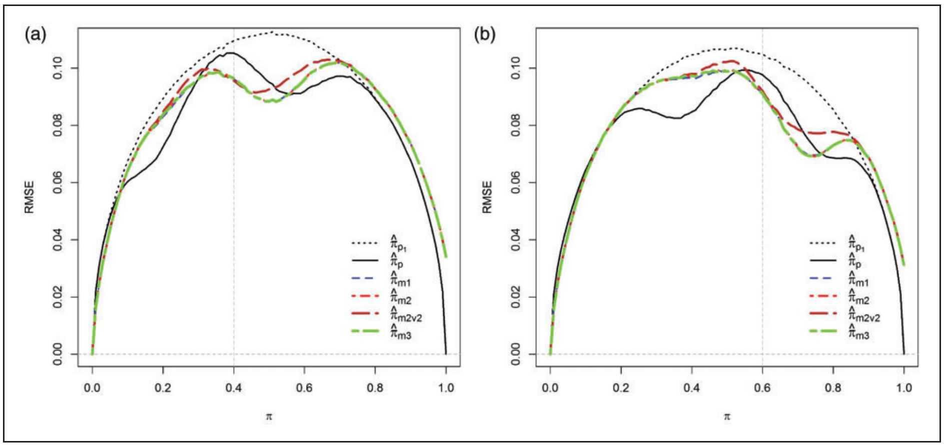


Figure 3. RMSE of estimators. Design 1 is defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and 2 by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. For each value of π, 50,000 trials were simulated. The vertical line represents $\pi = \pi_1$.

Table 4. Performance measures of estimator under H_1 (i.e. $\pi = \pi_1$).

	True response rate: $\pi = \pi_1$				
	$\hat{\pi}_p$	$\hat{\pi}_{m1}$	$\hat{\pi}_{m2}$	$\hat{\pi}_{m2v2}$	$\hat{\pi}_{m3}$
Mean	0.4126	0.3974	0.3966	0.4008	0.3968
Med.	0.4068	0.3978	0.3897	0.3975	0.3887
M.Bias	0.0126	−0.0026	−0.0034	0.0008	−0.0032
RMSE	0.1052	0.0962	0.0965	0.0956	0.0964
1*st* Q.	0.3559	0.3385	0.3414	0.3486	0.3408
3*rd* Q.	0.5000	0.4648	0.4677	0.4682	0.4677
Cov.P		0.9785	0.9785	0.9785	0.9785
M.LB		0.2685	0.2678	0.2660	0.2681
Mean	0.6057	0.5941	0.5933	0.5991	0.5935
Med.	0.6027	0.5983	0.5915	0.5976	0.5904
M.Bias	0.0057	−0.0059	−0.0067	−0.0009	−0.0065
RMSE	0.0976	0.0911	0.0912	0.0920	0.0911
1*st* Q.	0.5556	0.5502	0.5471	0.5524	0.5483
3*rd* Q.	0.6575	0.6422	0.6383	0.6599	0.6400
Cov.P		0.9742	0.9742	0.9742	0.9742
M.LB		0.4723	0.4716	0.4677	0.4718

Note: The measures are the mean, median (Med.), mean bias (M.Bias), RMSE, first and third quartiles (Q.), and the coverage probability (Cov.P) and mean of lower bound (M.LB) of the one-sided $(1-\alpha)100\%$ confidence interval. The first group of rows are for the design 1, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and the other for 2 $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. A total of 50,000 trials were simulated for each design.

Table 5. Performance measures of estimator under $\pi = \pi_1 + 0.1$.

	True response rate: $\pi = \pi_1 + 0.1$				
	$\hat{\pi}_p$	$\hat{\pi}_{m1}$	$\hat{\pi}_{m2}$	$\hat{\pi}_{m2v2}$	$\hat{\pi}_{m3}$
Mean	0.5265	0.4971	0.4970	0.4931	0.4971
Med.	0.5085	0.4754	0.4754	0.4754	0.4754
M.Bias	0.0265	−0.0029	−0.0030	−0.0069	−0.0029
RMSE	0.0950	0.0885	0.0884	0.0926	0.0883
1*st* Q.	0.4746	0.4434	0.4410	0.4283	0.4425
3*rd* Q.	0.6000	0.5737	0.5737	0.5737	0.5737
Cov.P		0.9785	0.9785	0.9785	0.9785
M.LB		0.3369	0.3353	0.3319	0.3369
Mean	0.7248	0.6978	0.6980	0.6925	0.6980
Med.	0.7273	0.6980	0.6994	0.6994	0.6982
M.Bias	0.0248	−0.0022	−0.0020	−0.0075	−0.0020
RMSE	0.0808	0.0726	0.0722	0.0786	0.0721
1*st* Q.	0.6761	0.6522	0.6502	0.6436	0.6509
3*rd* Q.	0.7727	0.7462	0.7462	0.7462	0.7462
Cov.P		0.9792	0.9792	0.9792	0.9792
M.LB		0.5513	0.5488	0.5416	0.5515

Note: The measures are the mean, median (Med.), mean bias (M.Bias), RMSE, first and third quartiles (Q.), and the coverage probability (Cov.P) and mean of lower bound (M.LB) of the one-sided $(1-\alpha)100\%$ confidence interval. The first group of rows are for the design 1, $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and the other for 2 $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. A total of 50,000 trials were simulated for each design.

approaches 0. Among the proposed estimators, $\hat{\pi}_{m2v2}$ behaves slightly different, as π moves from π_0 to 1, it is the first to attain mean bias that is closest to zero, which happens just after π_1. Regarding RMSE (see Figure 3), the proposed estimators also outperform $\hat{\pi}_p$ for values of π around π_1. Their RMSE is also lower than that of $\hat{\pi}_{p1}$.

From Table 4 we see that under H_1 the simulation mean and median are similar, and they are relatively lower in the proposed estimator as compared to $\hat{\pi}_p$. The one-sided CIs have good coverage probabilities for all the proposed methods, with values that are not less than the nominal level (95%). The mean of the lower bound of CIs are similar across the proposed methods. The simulation done assuming $\pi = \pi_1 + 0.1$ (Table 5) showed similar results.

10 Discussion

In this paper, we have discussed and proposed sample space orderings for oncology Phase II adaptive two-stage designs with binary endpoint. Overall *p*-value and point and interval estimation were derived from these sample space orderings. Although for some values of true response probability our methods do not show improvement over the fixed sample MLE, they are preferable because they consistently outperform the MLE when the true response probability is in the neighbourhood of values that are equal to or greater than the response probability under the alternative hypothesis. It is in this region where the estimation becomes particularly important since the null hypothesis would likely have been rejected and the treatment effect estimate needed to plan later Phase III trials. In this region, the MLE shows high positive bias and higher RMSE while our methods are either unbiased or negatively biased with smaller RMSE.

In general, as opposed to the fixed sample MLE, our proposed methods do not overestimate the response probability. This is seen by the fact that for values of true response rate raging from 0 to 1, they are either unbiased or negatively biased. Overestimation of treatment effect in Phase II trials has been acknowledged in the literature as one of the reasons for high failure rate of drugs in Phase III. For example, Kirby et al.[52] showed an evidence that supports the need to, and proposed methods to, discount the Phase II estimate of treatment effect when it is used to plan Phase III trial sample size.

Although our methods were built on top of Englert and Kieser[8] design, they can easily be extended to other similar designs with pre-specified adaptation rules. One example are the adaptive designs by Shan et al.[9,10] Further, some of our proposed approaches (methods 2 and 3) remain valid even if the adaptation rule is not pre-specified, meaning that they can be applied for flexible designs. Methods 2 and 3 can be extended to handle non-binary outcomes, while with the Method 1, that uses exact probability calculation, an extension would be difficult.

Acknowledgement

We would like to thank the two anonymous reviewers for their valuable comments and inputs.

Declaration of conflicting interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567.

References

1. Schultz JR, Nichol FR, Elfring GL, et al. Multiple-stage procedures for drug screening. *Biometrics* 1973; **29**: 293–300.
2. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* 1989; **10**: 1–10.
3. Lin Y and Shih WJ. Adaptive two-stage designs for single-arm phase IIA cancer clinical trials. *Biometrics* 2004; **60**: 482–490.
4. Banerjee A and Tsiatis AA. Adaptive two-stage designs in phase II clinical trials. *Stat Med* 2006; **25**: 3382–3395.
5. Englert S and Kieser M. Adaptive designs for single-arm phase II trials in oncology. *Pharm Stat* 2012; **11**: 241–249.
6. Englert S and Kieser M. Improving the flexibility and efficiency of phase II designs for oncology trials. *Biometrics* 2012; **68**: 886–892, http://dx.doi.org/10.1111/j.1541-0420.2011.01720.x
7. Jin H and Wei Z. A new adaptive design based on Simon's two-stage optimal design for phase II clinical trials. *Contemp Clin Trials* 2012; **33**: 1255–1260.
8. Englert S and Kieser M. Optimal adaptive two-stage designs for phase II cancer clinical trials. *Biom J* 2013; **55**: 955–968.
9. Shan G, Wilding GE, Hutson AD, et al. Optimal adaptive two-stage designs for early phase II clinical trials. *Stat Med* 2016; **35**: 1257–1266, http://dx.doi.org/10.1002/sim.6794. Sim.6794

10. Shan G, Zhang H and Jiang T. Minimax and admissible adaptive two-stage designs in phase II clinical trials. *BMC Med Res Methodol* 2016; **16**: 90, http://dx.doi.org/10.1186/s12874-016-0194-3
11. Chang MN, Wieand HS and Chang VT. The bias of the sample proportion following a group sequential phase II clinical trial. *Stat Med* 1989; **8**: 563–570.
12. Jung SH, Lee T, Kim K, et al. Admissible two-stage designs for phase II cancer clinical trials. *Stat Med* 2004; **23**: 561–569, http://dx.doi.org/10.1002/sim.1600
13. Guo HY and Liu A. A simple and efficient bias-reduced estimator of response probability following a group sequential phase II trial. *J Biopharm Stat* 2005; **15**: 773–781, http://dx.doi.org/10.1081/BIP-200067771
14. Koyama T and Chen H. Proper inference from simon's two-stage designs. *Stat Med* 2008; **27**: 3145–3154, http://dx.doi.org/10.1002/sim.3123
15. Tsai WY, Chi Y and Chen CM. Interval estimation of binomial proportion in clinical trials with a two-stage design. *Stat Med* 2008; **27**: 15–35, http://dx.doi.org/10.1002/sim.2930
16. Pepe MS, Feng Z, Longton G, et al. Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility. *Stat Med* 2009; **28**: 762–779, http://dx.doi.org/10.1002/sim.3506
17. Li Q. An mse-reduced estimator for the response proportion in a two-stage clinical trial. *Pharm Stat* 2011; **10**: 277–279, http://dx.doi.org/10.1002/pst.414
18. Jovic G and Whitehead J. An exact method for analysis following a two-stage phase II cancer clinical trial. *Stat Med* 2010; **29**: 3118–3125, http://dx.doi.org/10.1002/sim.3837
19. Jung SH and Kim KM. On the estimation of the binomial probability in multistage clinical trials. *Stat Med* 2004; **23**: 881–896, http://dx.doi.org/10.1002/sim.1653
20. Coburger S and Wassmer G. Sample size reassessment in adaptive clinical trials using a bias corrected estimate. *Biom J* 2003; **45**: 812–825, http://dx.doi.org/10.1002/bimj.200390051
21. Cheng Y and Shen Y. Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics* 2004; **60**: 910–918, http://dx.doi.org/10.1111/j.0006-341X.2004.00246.x
22. Posch M, Koenig F, Branson M, et al. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Stat Med* 2005; **24**: 3697–3714, http://dx.doi.org/10.1002/sim.2389
23. Stallard N and Todd S. Point estimates and confidence regions for sequential trials involving selection. *J Stat Plan Inference* 2005; **135**: 402–419, http://dx.doi.org/10.1016/j.jspi.2004.05.006
24. Brannath W, König F and Bauer P. Estimation in flexible two stage designs. *Stat Med* 2006; **25**: 3366–3381, http://dx.doi.org/10.1002/sim.2258
25. Bowden J and Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biom J* 2008; **50**: 515–527, http://dx.doi.org/10.1002/bimj.200810442
26. Bebu I, Luta G and Dragalin V. Likelihood inference for a two-stage design with treatment selection. *Biom J* 2010; **52**: 811–822, http://dx.doi.org/10.1002/bimj.200900170
27. Luo X, Li M, Shih WJ, et al. Estimation of treatment effect following a clinical trial with adaptive design. *J Biopharm Stat* 2012; **22**: 700–718, http://dx.doi.org/10.1080/10543406.2012.676534
28. Bebu I, Dragalin V and Luta G. Confidence intervals for confirmatory adaptive two-stage designs with treatment selection. *Biomed J* 2013; **55**: 294–309, https://dx.doi.org/10.1002/bimj.201200053
29. Carreras M and Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Stat Med* 2013; **32**: 1677–1690.
30. Gao P, Liu L and Mehta C. Exact inference for adaptive group sequential designs. *Stat Med* 2013; **32**: 3991–4005, http://dx.doi.org/10.1002/sim.5847
31. Bowden J, Brannath W and Glimm E. Empirical bayes estimation of the selected treatment mean for two-stage drop-the-loser trials: a meta-analytic approach. *Stat Med* 2014; **33**: 388–400, http://dx.doi.org/10.1002/sim.5920
32. Bowden J and Glimm E. Conditionally unbiased and near unbiased estimation of the selected treatment mean for multistage drop-the-losers trials. *Biomed J* 2014; **56**: 332–349, http://dx.doi.org/10.1002/bimj.201200245
33. Bowden J and Trippa L. Unbiased estimation for response adaptive clinical trials. *Stat Meth Med Res* 2017; **26**: 2376–2388.
34. Kimani PK, Todd S and Stallard N. Estimation after subpopulation selection in adaptive seamless trials. *Stat Med* 2015; **34**: 2581–2601, http://dx.doi.org/10.1002/sim.6506
35. Broberg P and Miller F. Conditional estimation in two-stage adaptive designs. *Biometrics* 2017; **73**: 895–904.
36. Kunzmann K and Kieser M. Point estimation and p-values in phase II adaptive two-stage designs with a binary endpoint. *Stat Med* 2017; **36**: 971–984, http://dx.doi.org/10.1002/sim.7200. Sim.7200
37. Bezanson J, Karpinski S, Shah VB, et al. Julia: a fast dynamic language for technical computing. *arXiv preprint arXiv:12095145* 2012.
38. Lubin M and Dunning I. Computing in operations research using julia. *INFORMS J Comput* 2015; **27**: 238–248.
39. Gurobi Optimization Inc. Gurobi optimizer reference manual, http://www.gurobi.com (2015).
40. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017, https://www.R-project.org/
41. Armitage P. Restricted sequential procedures. *Biometrika* 1957; **44**: 9–26.
42. Siegmund D. Estimation following sequential tests. *Biometrika* 1978; **65**: 341–349.

43. Fairbanks K and Madsen R. P values for test using a repeated significance test design. *Biometrika* 1982; **69**: 69–74.
44. Tsiatis AA, Rosner GL and Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics* 1984; **40**: 797–803.
45. Jennison C and Turnbull BW. *Group sequential methods with applications to clinical trials*. Boca Raton, Florida: Chapman-Hall/CRC, 2000.
46. Wassmer G and Brannath W. *Group sequential and confirmatory adaptive designs in clinical trials*. Springer Series in Pharmaceutical Statistics, Springer International Publishing, 2016, https://books.google.de/books?id=lhWJjwEACAAJ.
47. Chang MN and O'Brien PC. Confidence intervals following group sequential tests. *Control Clin Trials* 1986; **7**: 18–26.
48. Chang MN. Confidence intervals for a normal mean following a group sequential test. *Biometrics* 1989; **45**: 247–254.
49. Rosner GL and Tsiatis AA. Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* 1988; **75**: 723–729.
50. Emerson SS and Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990; **77**: 875–892, http://www.jstor.org/stable/2337110
51. Lehmacher W and Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**: 1286–1290.
52. Kirby S, Burke J, Chuang-Stein C, et al. Discounting phase 2 results when planning phase 3 clinical trials. *Pharm Stat* 2012; **11**: 373–385.

Appendix 1. Details of Method 3

Let's first show how $p_{1b}(x_1)$ is obtained

$$\begin{aligned}
A(p_{1b}(x_1)) &= D(x_1)\\
&\Leftrightarrow 1-\Phi\left\{\frac{\Phi^{-1}(1-c)-w_1\Phi^{-1}[1-p_{1b}(x_1)]}{w_2}\right\}=D(x_1)\\
&\Leftrightarrow \frac{\Phi^{-1}(1-c)-w_1\Phi^{-1}[1-p_{1b}(x_1)]}{w_2}=\Phi^{-1}[1-D(x_1)]\\
&\Leftrightarrow \Phi^{-1}[1-p_{1b}(x_1)]=\frac{\Phi^{-1}(1-c)-w_2\Phi^{-1}[1-D(x_1)]}{w_1}\\
&\Leftrightarrow 1-p_{1b}(x_1)=\Phi\left\{\frac{\Phi^{-1}(1-c)-w_2\Phi^{-1}[1-D(x_1)]}{w_1}\right\}\\
&\Leftrightarrow p_{1b}(x_1)=1-\Phi\left\{\frac{\Phi^{-1}(1-c)-w_2\Phi^{-1}[1-D(x_1)]}{w_1}\right\}
\end{aligned}$$

We have that

$$\begin{aligned}
\delta(x_1',x_2') &\geq \delta(x_1,x_2)\\
&\Leftrightarrow \delta\big[p_{1b}(x_1'),p_2(x_2')\big]\geq\delta[p_{1b}(x_1),p_2(x_2)]\\
&\Leftrightarrow 1-C(p_{1b}',p_2')\geq 1-C(p_{1b},p_2)\\
&\Leftrightarrow C(p_{1b}',p_2')\leq C(p_{1b},p_2)\\
&\Leftrightarrow 1-\Phi\big[w_1'\Phi^{-1}\big(1-p_{1b}'\big)+w_2'\Phi^{-1}\big(1-p_2'\big)\big]\\
&\leq 1-\Phi\big[w_1\Phi^{-1}(1-p_{1b})+w_2\Phi^{-1}(1-p_2)\big]\\
&\Leftrightarrow \Phi\big[w_1'\Phi^{-1}\big(1-p_{1b}'\big)+w_2'\Phi^{-1}\big(1-p_2'\big)\big]\\
&\geq \Phi\big[w_1\Phi^{-1}(1-p_{1b})+w_2\Phi^{-1}(1-p_2)\big]\\
&\Leftrightarrow w_1'\Phi^{-1}\big(1-p_{1b}'\big)+w_2'\Phi^{-1}\big(1-p_2'\big)\\
&\geq w_1\Phi^{-1}(1-p_{1b})+w_2\Phi^{-1}(1-p_2)\\
&\Leftrightarrow \Phi^{-1}\big(1-p_2'\big)\\
&\geq \frac{w_1\Phi^{-1}(1-p_{1b})+w_2\Phi^{-1}(1-p_2)-w_1'\Phi^{-1}\big(1-p_{1b}'\big)}{w_2'}\Leftrightarrow 1-p_2'\\
&\geq \Phi\left[\frac{w_1\Phi^{-1}(1-p_{1b})+w_2\Phi^{-1}(1-p_2)-w_1'\Phi^{-1}\big(1-p_{1b}'\big)}{w_2'}\right]\Leftrightarrow p_2'\\
&\leq 1-\Phi\left[\frac{w_1\Phi^{-1}(1-p_{1b})+w_2\Phi^{-1}(1-p_2)-w_1'\Phi^{-1}\big(1-p_{1b}'\big)}{w_2'}\right]
\end{aligned}$$

Note that some of steps above are only valid because Φ is a continuous and monotone increasing function.

Details on point and interval estimation

Let π_0 be the response probability under the design's original null hypothesis H_0, and we denote by $\tilde{\pi}_0$ the response probability under all null hypotheses we consider when searching for point and interval estimates, with $0 \leq \tilde{\pi}_0 \leq 1$. Then the overall p-value to search for the estimates is defined as

$$Q(\tilde{\pi}_0) = \begin{cases} 1 - B(x_1 - 1, n_1, \tilde{\pi}_0) & \text{if } m = 1 \\ 1 - B(u_1 - 1, n_1, \tilde{\pi}_0) & \\ + \sum\limits_{x_1' = l_1 + 1}^{u_1 - 1} b(x_1', n_1, \tilde{\pi}_0) \, \mathrm{Pr}_{\tilde{\pi}_0}\left(\Delta \geq \delta | x_1'\right) & \text{if } m = 2 \end{cases}$$

where $\mathrm{Pr}_{\tilde{\pi}_0}(\Delta \geq \delta) = \mathrm{Pr}_{\tilde{\pi}_0}[\delta(X_1, X_2) \geq \delta(x_1, x_2)]$ is calculated for the different methods as follows. For the Method 1

$$\Pr_{\tilde{\pi}_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | X_1 = x_1'\right] = 1 - B\left[x - l(x_1) + l(x_1') - x_1' - 1, n_2(x_1'), \tilde{\pi}_0\right]$$

For the Method 2

$$\Pr_{\tilde{\pi}_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | X_1 = x_1'\right] \approx \left\langle \tilde{p}_2 - D(x_1) + D(x_1') \right\rangle_{[0,1]}$$

where $\tilde{p}_2 = 1 - B[x_2 - 1, n_2(x_1), \tilde{\pi}_0]$

For the Method 2v2

$$\Pr_{\tilde{\pi}_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | X_1 = x_1'\right] = 1 - B\left\{B_q\left[1 - p_2 + D(x_1) - D(x_1'), n_2(x_1'), \pi_0\right], n_2(x_1'), \tilde{\pi}_0\right\}$$

with $p_2 = 1 - B[x_2 - 1, n_2(x_1), \pi_0]$.

And for the Method 3

$$\Pr_{\tilde{\pi}_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2) | X_1 = x_1'\right] \approx 1 - \Phi\left[\frac{w_1 \Phi^{-1}(1 - p_{1b}) + w_2 \Phi^{-1}(1 - \tilde{p}_2) - w_1' \Phi^{-1}\left(1 - p_{1b}'\right)}{w_2'}\right]$$

where $\tilde{p}_2 = 1 - B[x_2 - 1, n_2(x_1), \tilde{\pi}_0]$

Appendix E

Second paper

Using Estimates from Adaptive Phase II Oncology Trials to Plan Phase III Trials

Arsénio Nhacolo^{*,1} and **Werner Brannath** [1]

[1] Competence Centre for Clinical Trials, Universität Bremen, Linzerstraße 4, 28359 Bremen, Germany

The clinical drug development is mainly done in three phases, Phase I, Phase II and Phase III. The knowledge gained in clinical trials of a particular phase is often used to plan trials of subsequent phases. That is the case with successful Phase II clinical trials in which, among others aspects, the effect size estimates are used to plan the sample size of the related Phase III trials. Due to small sample sizes, selections bias and other factors, Phase II estimates are often biased and imprecise, resulting in inadequately powered Phase III trials. We evaluated through simulation studies the consequences, in terms of power, of using the effect estimate from Phase II adaptive design trials to plan sample size of Phase III trials in oncology. In addition, we propose new approaches for adjusting Phase II estimates. We considered the recently proposed oncology Phase II two-stage single-arm adaptive designs with binary endpoint, in which the second stage sample size is a pre-defined function of the first stage's number of responses. We used the naïve and the recently proposed estimators for estimating the Phase II effect. Results showed that using naïve estimates lead to underpowered Phase III trials, while estimates that take into account the adaptiveness of the designs lead to power close to the target value. Our new adjustment approach seems to perform well for all estimation methods. It also showed that a relatively higher *discount* is necessary for naïve estimates.

Key words: Adaptive Design, Bias, Clinical Trials, Estimation, Bootstrap, Phase II, Phase III, Power, Sample Size

1 Introduction

Drug development is a lengthy and costly process that spans different phases, from pre-clinical studies to post-marketing surveillance trials. The accumulating knowledge gained from studies of a particular phase is often used to better inform decisions on conducting studies of subsequent phases. Confirmatory clinical trials done in Phase III are of paramount importance as they are intended to provide firm evidence to support drug approval for use. The decision to conduct these trials is mainly justified by positive results from clinical trials of Phase II. The effect size estimates from successful Phase II trials are often used to plan the sample size of the related Phase III trials. Due to various factors, including small sample sizes and selections bias, Phase II estimates are often biased and imprecise, resulting in Phase III trials that are not properly powered. Associated to this is the high failure rate of Phase III clinical trials, which is approximately 40% in general (De Martini, 2013) and 60% in oncology (Gan et al., 2012). Another issue that contribute to these high failure rates are the overly optimistic assumptions regarding treatment benefits that investigators tend to make when designing Phase III trials (Gan et al., 2012). This excessive optimism might in part result from the problem of over-estimation of the treatment effect in Phase II trials. Acknowledging these shortcomings, many authors (e.g., Kirby et al., 2012; Wang et al., 2006; Burke et al., 2014) have discussed and proposed approaches to make adjustments of estimates from Phase II trials when using them to plan confirmatory studies.

In this paper we evaluate through simulation studies the consequences, in terms of power, of using the effect estimate from oncology Phase II adaptive design trials to plan sample size of a related Phase III trial.

*Corresponding author: e-mail: anhacolo@uni-bremen.de

In addition, we propose and discuss new approaches to obtain multiplicative adjustment factors for Phase II estimates and/or Phase III sample sizes based on the Phase II data. We consider the recently proposed oncology Phase II two-stage single-arm adaptive designs with binary endpoint, in which the second stage sample size is a pre-defined function of the first stage's number of responses (successes). Examples of such designs are given in Englert & Kieser (2013) and Shan et al. (2016). Different estimators will be used. The naïve (fixed-sample) maximum likelihood estimator (MLE), which is more likely to be employed for adaptive design for which no specific estimator has been proposed, and the estimates recently proposed for such designs by Nhacolo & Brannath (2018). For simplicity, we consider two-arm Phase III RCTs also with binary endpoint. Although a survival endpoint is commonly used in oncology Phase III trials, there are some types of cancer for which the response rate is a suitable endpoint. The objective response rate (ORR), as defined by the Response Criteria in Solid Tumours (RECIST) guidelines (Eisenhauer et al., 2009), is the most commonly used binary endpoint in oncology trials. ORR has been used as the primary endpoint in 40% of advanced breast cancer Phase III trials published between January 1998 and December 2007 (Saad et al., 2010).

This paper is organized as follows. We first present a brief literature review of approaches on how to adjust Phase II effect estimates before employing them to plan Phase III sample size. We then propose new adjustment approaches. Then a summary of the trial designs and estimation methods used in simulations follows. Next we present the set-up of the simulation study and the results. The simulation study has two parts. The first one evaluates different Phase II estimators with respect to Phase III power, and the second part evaluates our proposed adjustment methods applied to different estimators also with respect to Phase III power. The paper ends with a summary and final discussion.

2 Dealing with bias and imprecision of Phase II estimates

Different approaches to deal with bias and imprecision of Phase II treatment effect estimates when planning Phase III sample sizes have been proposed in the literature (see Wang et al., 2006; De Martini, 2011a,b; Kirby et al., 2012; De Martini, 2013; Burke et al., 2014; Chuang-Stein & Kirby, 2017). Most of these approaches fall into the category of *conservative sample size estimation* (CSSE) strategies, with some following frequentist methods and others Bayesian. The frequentist CSSE strategies consist in using a conservative value, $\hat{\theta}_f$, of Phase II effect estimate, $\hat{\theta}$, to determine Phase III sample size. This can be achieved by subtracting a certain amount from $\hat{\theta}$ (Wang et al., 2006), e.g., one standard error, leading to $\hat{\theta}_f = \hat{\theta} - \mathrm{SE}(\hat{\theta})$, or by applying a discounting factor $f \in]0,1]$, resulting in $\hat{\theta}_f = \hat{\theta} \times f$ (Kirby et al., 2012). The Bayesian CSSE put a probability mass around the observed Phase II effect and computes the averaged success probability (SP) at a given sample size. Then the Phase III sample size estimate is the minimum sample size whose Bayesian SP exceeds a certain desired power. When many similar Phase II trials on the same therapy exist, meta-analytic approaches can also be used to better plan subsequent Phase III trials. For instance, in the context of randomized Phase II trials with binary endpoints, Burke et al. (2014) deemed meta-analysis using a Bayesian random effects logistic regression model to be the most appropriate. The model can predict the probability that the therapy will be truly effective in a new trial and that in a new trial with a given sample size, the 95% credible interval for the odds ratio will be entirely in favour of the therapy.

3 New approaches for adjusting Phase II estimates

As it will be seen in the simulations results below, although some estimates from Phase II designs suffer less from bias and imprecision, it is nearly always necessary to adjust them when they are used to calculate the Phase III sample size. Many approaches to make these adjustments exist, however their implementation in practice is cumbersome. This is due to the difficulties in establishing clear guidelines, for instance, on the adequate amount to *discount* from the estimates for the frequentist conservative sample size estimation

strategies or on the adequate choice of the conservative prior distribution for Bayesian strategies.
We propose new approaches that, given the observed Phase II data, estimate the multiplicative adjustment factor to be applied to Phase II treatment effect estimates before employing them to plan the sample size of the related Phase III clinical trials. Alternatively, the approaches do also estimate the adjustment factor to be applied to the Phase III sample size planned using unadjusted Phase II efficacy estimates. The proposed approaches are based on parametric bootstrapping.

3.1 Notation

Let θ be the true efficacy parameter, θ_0 its value under the null hypothesis (H_0) of no treatment effect and θ_1 its value under the alternative (H_1). Further let τ be an actual Phase II trial testing the hypothesis $H_0 : \theta \leq \theta_0$ versus $H_0 : \theta \geq \theta_1$, and Y the corresponding observed data, drawn from a parametric distribution $\mathcal{F}(\theta)$. Denote the estimate of θ, given Y, by $\hat{\theta}$. Let $\tau^\star$ be a simulated trial following the same design as τ but assuming $\hat{\theta}$ to be the true efficacy parameter, $Y^\star$ the resulting data, which is drawn from the distribution $\mathcal{F}(\hat{\theta})$, and $\theta^\star$ the corresponding estimate (of $\hat{\theta}$). Assume that the type I and II error rates of interest for the subsequent Phase III trial are α and β, respectively. Denote the Phase III sample size assuming $\hat{\theta}$ as the efficacy under H_1 by $\hat{n} = \hat{n}(\hat{\theta}, \alpha, \beta)$, and assuming $\theta^\star$ by $n^\star = n^\star(\theta^\star, \alpha, \beta)$. Let f be the multiplicative adjustment factor to be applied to the Phase II effect estimate, and ρ the multiplicative factor to be applied to the Phase III sample size.

3.2 Method 1

This method estimates both f and ρ. Assume that a Phase II trial τ was conducted and that H_0 was rejected at the end, leading to the decision to proceed for further testing of the treatment in a Phase III trial. We calculate $\hat{\theta}$ from the observed data, and based on it we simulate Phase II trials $\tau_i^\star$, $i = 1, \ldots, k$. Let J be the index set of all the simulated trials in H_0 was rejected, and $|J|$ its cardinality ($|J| \leq k$) . For each $\tau_j^\star$, $j \in J$, we calculate $\theta_j^\star$ and the respective Phase III sample size, $n_j^\star = n^\star(\theta_j^\star, \alpha, \beta)$. Then the individual estimates of the multiplicative adjustment factors for the effect size and sample size, f_j and ρ_j, are calculated as $f_j = \hat{\theta}/\theta_j^\star$ and $\rho_j = \hat{n}/n_j^\star$, where $\hat{n} = \hat{n}(\hat{\theta}, \alpha, \beta)$. We take the average values as the final estimates, i.e., $f = \frac{1}{|J|} \sum_{j \in J} f_j$ and $\rho = \frac{1}{|J|} \sum_{j \in J} \rho_j$. Hence, the adjusted efficacy to be used in planing Phase III trials is $\theta_f = \hat{\theta} \times f$. Alternatively, the adjusted sample size $n_\rho = \hat{n} \times \rho$ could be used. Adjustments using f and ρ lead, in general, to different sample sizes and, as consequence, to differences in Phase III power. This is because the sample size is generally a non-linear function of the effect size.

3.3 Method 2

This is an alternative method to estimate the sample size adjustment factor ρ. Let $\mathrm{pwr}_j^\star = \mathrm{pwr}(n_j^\star, \hat{\theta}, \alpha)$ be the power that a Phase III trial would attain, with sample size $n_j^\star$ assuming that the true efficacy (θ) is equal to the observed estimate ($\hat{\theta}$). The sample size multiplicative adjustment factor is calculated such that the expected value of pwr* is equal to the desired power assuming (as an approximation) that $\hat{\theta}$ and θ are equal, i.e.,

$$\rho = \left\{ \tilde{\rho} | E_{\hat{\theta}} \left[\mathrm{pwr}(\tilde{\rho} n_j^\star, \hat{\theta}, \alpha) \right] = 1 - \beta \right\}.$$

We employ numerical root finding to get the adequate value of ρ.

4 Trial designs and estimation methods

We give a brief summary of designs and estimation methods that we use in the simulation study below. For the Phase II, we consider the adaptive single-arm two-stage designs with a binary endpoint proposed by Englert & Kieser (2013) for oncology trials. These designs test, at type I error rate α and type II error rate β, the null (H_0) versus the alternative (H_1) hypotheses about the response rate π, $H_0 : \pi \leq \pi_0$ vs $H_1 : \pi \geq \pi_1$, where π_0 is the maximum response rate considered to be uninteresting and π_1 is the minimum desirable response rate, with $\pi_1 > \pi_0$. The designs are defined by static elements in the first stage, namely the sample size n_1, and the futility and efficacy boundaries, l_1 and u_1 ($u_1 > l_1$), and in the second stage by elements that vary depending on the number of responses observed in the first stage (x_1), namely the sample size, $n_2(x_1)$, the conditional error function, $D(x_1)$, and the corresponding decision boundary, $l(x_1)$. With these designs, trials are stopped at the first stage with no rejection of H_0 if $x_1 \leq l_1$ or with rejection of H_0 if $x_1 \geq u_1$. Otherwise the trial proceeds to the second stage, at which H_0 is rejected if $p_2 < D(x_1)$ or, equivalently, $x > l(x_1)$, where p_2 is the second stage p-value and x is the total number of responses.

For Phase III we consider a single-stage randomized parallel-group trial design with binary endpoint, similar to that described by Halabi (2008). The design tests the null hypothesis that the response rate in the control and treatment groups, π_c and π_t, are equal, i.e., $H_0 : \pi_c = \pi_t$, against $H_1 : \pi_c \neq \pi_t$.

We used different estimators to get the treatment efficacy estimates from Phase II trials, the naïve maximum likelihood estimator (MLE), which ignores the adaptive nature of the designs, and the estimators proposed by Nhacolo & Brannath (2018). The naïve MLE is defined as

$$\hat{\pi}_{nml} = x/n.$$

Let (m, x_1, x) be the outcome of atrial that stopped at stage m with first stage's and total number of responses x_1 an x. To propose their estimators, Nhacolo & Brannath (2018) modified the classical stage-wise sample space ordering to take into account the design's adaptation rule by defining a function $\delta(x_1, x_2)$ that orders the outcome space in a way that is consistent with the rejection boundary of the trial design. They use three different methods to define $\delta(x_1, x_2)$. In the first method δ is defined using x as $\delta(x_1, x_2) = x - l(x_1)$, in the second method defined using the second stage p-value as $\delta(x_1, x_2) = D(x_1) - p_2(x_2)$ and, finally, in the third method defined as $\delta(x_1, x_2) = 1 - C(p_{1_b}, p_2)$, where C is the weighted inverse normal combination function represented as $C(p_1, p_2) = 1 - \Phi\left[w_1\Phi^{-1}(1 - p_1) + w_2\Phi^{-1}(1 - p_2)\right]$, with

$$w_1 = \sqrt{\frac{n_1}{n_1 + n_2(x_1)}}, w_2 = \sqrt{\frac{n_2(x_1)}{n_1 + n_2(x_1)}}$$

and

$$p_{1_b}(x_1) = 1 - \Phi\left\{\frac{\Phi^{-1}(1 - c) - w_2\Phi^{-1}\left[1 - D(x_1)\right]}{w_1}\right\}.$$

Based on these sample space ordering, they derive the overall p-value Q, calculated as

$$Q = \begin{cases} 1 - B(x_1 - 1, n_1, \tilde{\pi}_0) & \text{if } m = 1 \\ 1 - B(u_1 - 1, n_1, \tilde{\pi}_0) + \sum\limits_{X_1 = l_1 + 1}^{u_1 - 1} b(X_1, n_1, \tilde{\pi}_0) \Pr_{\tilde{\pi}_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2)\right] & \text{if } m = 2 \end{cases}$$

where $B(x, n, \pi)$ and $b(x, n, \pi)$ are the binomial cumulative distribution function and probability mass function with x successes, n trials and success probability π. Finally, they define the point estimate as

$$\hat{\pi}_m = \{\tilde{\pi}_0 : Q(\tilde{\pi}_0) = \underset{\tilde{\pi}_0}{\Pr}\left((M, X_1, X) \succeq (m, x_1, x)\right) = 0.5\}.$$

We follow the same notation used by Nhacolo & Brannath (2018) and denote the estimated response probability by $\hat{\pi}_{m1}$ for δ defined in terms of number of responses and rejection boundary, $\hat{\pi}_{m2}$ and $\hat{\pi}_{m2v2}$ for for δ defined in terms of second stage p-value and conditional error function (here two different approaches are used to calculate $\Pr_{\tilde{\pi}_0}\left[\delta(X_1, X_2) \geq \delta(x_1, x_2)\right]$, hence the two notations), and $\hat{\pi}_{m3}$ for δ defined by the inverse normal combination function.

5 Simulation study

The simulations study was done in two parts. The first part aimed at assessing the consequences of using the estimates from oncology Phase II adaptive designs to plan the sample size of the subsequent Phase III trials, with respect to the statistical power. The second part evaluates the performance of our proposed adjustment methods, also with respect to the power of Phase III trials.

The simulation procedure for the first part is as follows. From the Phase II designs described above, we select one to test $H_0 : \pi \leq \pi_0$ versus $H_1 : \pi \geq \pi_1$ at determined α and β, and simulate K trials assuming a specific true response probability π. We discard simulated trials that failed to reject H_0, and from the J remaining trials (i.e., trials in which H_0 was rejected) we get the individual estimates of π, denoted $\hat{\pi}_j$, $j = 1, \ldots, J$. We pick a Phase III design described above, testing $H_0 : \pi_c = \pi_t$ versus $H_1 : \pi_c \neq \pi_t$ at the desired type I and type II error rates α' and β', respectively, with $\pi_c = \pi_0$. Then we calculate the required sample size, N_j, to detect the effect size of magnitude $\hat{\pi}_j - \pi_c$ with power of $1 - \beta'$. Using N_j, we calculate what would be the attained power to detect the true effect size, $\pi - \pi_c$. All the estimators mentioned in the previous section were used to obtain $\hat{\pi}_j$. In addition to using the unadjusted estimates to calculate N_j, different values of the multiplicative adjustment factor, f, ($f \in [0, 1]$) proposed by Kirby et al. (2012) were used to obtain the adjusted estimates $\hat{\pi}_{fj} = \hat{\pi}_j \times f$. Note that unlike Kirby et al. (2012), we do not define a launch criterion based on a threshold of $\hat{\pi}_f$, instead we assume that a Phase III trial is launched whenever the null hypothesis is rejected in the Phase II trial. Note also that since we exclude the unsuccessful Phase II trials, the Phase III power we are calculating is conditional on rejection of H_0 in Phase II. We do so because we think that in practice, for the planning of Phase III studies, only positive Phase II trials are likely to be used.

We use the two-sample test for proportion described by Ahn et al. (2014) for calculating the power and sample size for the Phase III trials. The power is approximated by

$$\Phi\left(\frac{\pi_t - \pi_c}{\sqrt{\pi_t(1-\pi_t)/N_t + \pi_c(1-\pi_c)/N_c}} - z_{1-\alpha/2}\right),$$

and the sample size $N = N_c + N_t$ needed to achieve a power of $1 - \beta$ obtained by solving the equation

$$\frac{\pi_t - \pi_c}{\sqrt{\pi_t(1-\pi_t)/N_t + \pi_c(1-\pi_c)/N_c}} - z_{1-\alpha/2} = z_{1-\beta},$$

where Φ and z_u are the standard normal cumulative distribution function and u-quantile, and N_t and N_c are the sample sizes for the treatment and control groups, respectively. In the simulations, we assume equal size groups, hence

$$N_t = N_c = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\pi_t - \pi_c\right)^2}\left[\pi_t(1-\pi_t) + \pi_c(1-\pi_c)\right].$$

Two different Phase II designs were used, one design for $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$, and the other for $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 22)$. We varied the true response rate π from $\pi_0 + 0.1$ to $\pi_1 + 0.3$ by increments of 0.01. The assumed Phase III type I error rate and power are 5% and 90%, and the retention factor f varied from 0.5 to 1 by increments of 0.01. Note that $f = 1$ means no effect is retained, i.e., the original estimate is used. For each scenario 50000 Phase II trials were simulated.

For the second part, which aimed at examining the extent to which the our proposed adjustment methods lead to adequate power of Phase III trials, we used the same designs (with the same type I error and power), and the same methods for estimating the Phase II effect and Phase III sample size and power, as in the first simulations above. The simulation procedure was as follows. Simulate Phase II trials assuming a specific θ. For each successful trial (i.e., with H_0 rejected) apply our proposed adjustment methods to get the Phase III sample size. Then, given this sample size and θ, calculate the power of the Phase III trial. The Phase II design has a binary endpoint, therefore, the distribution $\mathcal{F}$ described in Section 3 is binomial with parameter π. The Phase II designs are simulated under the alternative hypothesis, i.e., $\pi = \pi_1$. For each scenario, 5000 trials were simulated and for each successful trial our adjustment methods generated another 5000 bootstrap samples. Note that due to the independence of the phase II and Phase III data, type I error control is out of question and hence need not be investigated in the simulations.

All the simulations and computation were done using the statistical programming language `R` (R Core Team, 2017).

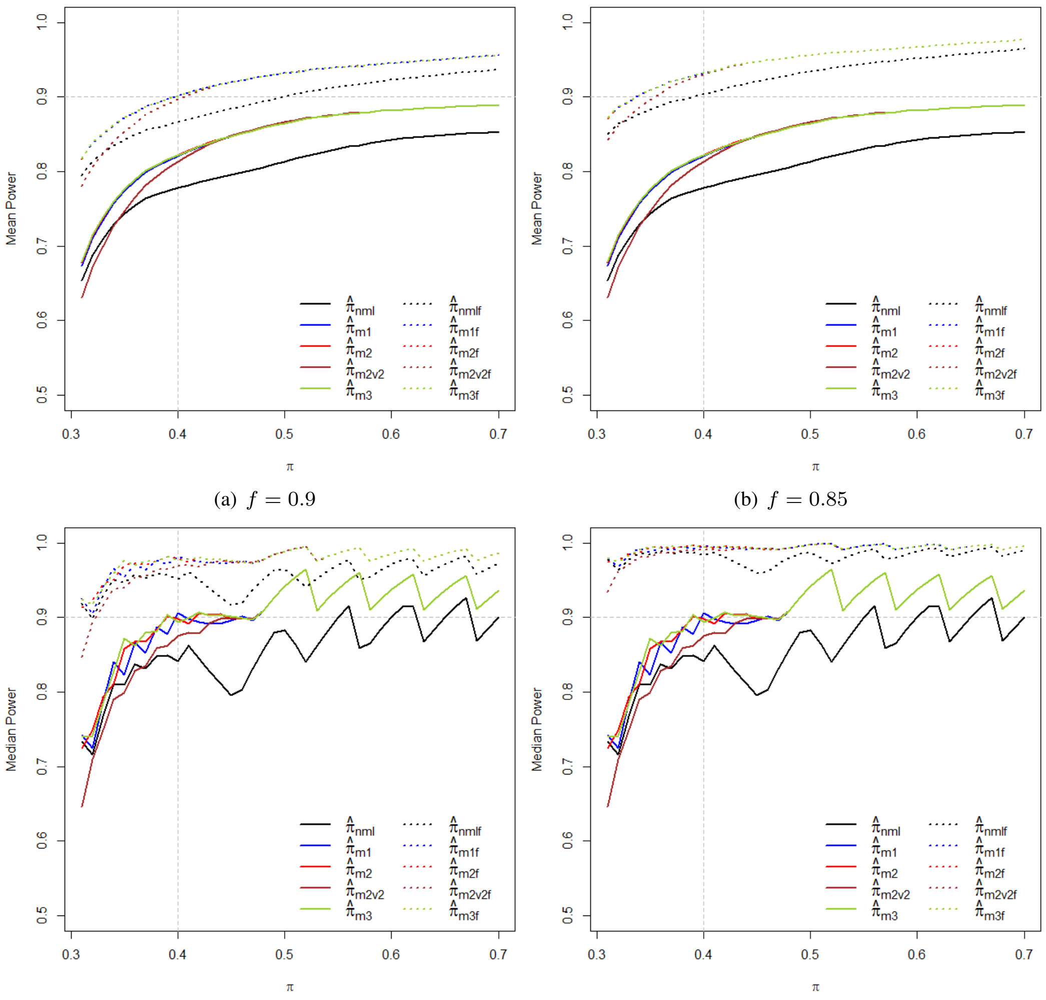
6 Results

The results of the first part of the simulation study are shown in the Tables 1 and 2 and Figures 1 and 2. These results are regarding the power attained by Phase III trials, the sample size of which was estimated using unadjusted and adjusted effect estimates from adaptive Phase II trials. It can be seen that the effect estimates from the methods that take into account the adaptive nature of the designs ($\hat{\pi}_{m1}$, $\hat{\pi}_{m2}$, $\hat{\pi}_{m2v2}$ and $\hat{\pi}_{m3}$) yield, in general, better power as compared to the naïve estimate ($\hat{\pi}_{nml}$). For instance, for the trials simulated under H_1 in the Table 1, when no adjustment is applied (i.e., $f = 1$), $\hat{\pi}_{nml}$ yielded a mean power of 77.8% while the power from the other effect estimates was around 82%, although still lower than the target power (90%). The median power is higher than the mean in all cases, with the naïve estimate yielding 84.1% and the others values between 87.6% and 90%. In order to attain a mean power that is equal to the target, an adjustment factor of $f = 0.85$ is necessary for the naïve estimate, while for the other estimates $f = 0.9$ is sufficient. When the values of the true response rate (π) are varied (Figures 1 and 2), results show that all the estimates yield under-powered Phase III trials for π less than π_1. But as π increases, the attained power using the adaptive estimates becomes higher than that of the naïve estimate. As π gets higher than π_1, the power among the adaptive estimates becomes more homogeneous.

Table 1 Mean and median power of Phase III trials planned using Phase II design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$. The target power of the Phase III trial is of 90%.

		$f = 1$		$f = 0.9$		$f = 0.85$	
π	$\hat{\pi}$	Mean	Median	Mean	Median	Mean	Median
π_1	$\hat{\pi}_{nml}$	0.7779	0.8414	0.8660	0.9515	0.9035	0.9828
	$\hat{\pi}_{m1}$	0.8207	0.9060	0.9009	0.9818	0.9321	0.9958
	$\hat{\pi}_{m2}$	0.8221	0.8986	0.9013	0.9789	0.9322	0.9949
	$\hat{\pi}_{m2v2}$	0.8131	0.8755	0.8969	0.9690	0.9300	0.9910
	$\hat{\pi}_{m3}$	0.8219	0.8936	0.9013	0.9769	0.9322	0.9941
$\pi_1 + 0.1$	$\hat{\pi}_{nml}$	0.8131	0.8834	0.9004	0.9646	0.9340	0.9863
	$\hat{\pi}_{m1}$	0.8648	0.9416	0.9315	0.9887	0.9556	0.9970
	$\hat{\pi}_{m2}$	0.8648	0.9416	0.9316	0.9887	0.9556	0.9970
	$\hat{\pi}_{m2v2}$	0.8660	0.9416	0.9317	0.9887	0.9556	0.9970
	$\hat{\pi}_{m3}$	0.8648	0.9416	0.9316	0.9887	0.9556	0.9970

The results from the second part of the simulation study, regarding our proposed adjustment methods, are shown in the Table 3. As expected from the pattern seen in the previous simulation results, the naïve MLE yields a lower value for the effect estimate adjustment factor f and, conversely, a higher value of sample size adjustment factor ρ as compared to the estimates that take into account the adaptiveness of the designs. This means that, to achieve the power goals, the naïve MLE requires that a higher amount of effect



(a) $f = 0.9$

(b) $f = 0.85$

(c) $f = 0.9$

(d) $f = 0.85$

Figure 1 Mean and median power of Phase III trials planned using Phase II design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$. The solid lines are for the unadjusted effect estimates, and the dashed lines for the adjusted ones (i.e., estimates multiplied by the retention factor f). For each value of π 50000 trials were simulated. The vertical line represents $\pi = \pi_1$ and the horizontal one represents the target power (90%).

is discounted as compared to the others. Regarding the mean power, our methods show improvements, with the Method 2 being the best. For instance, in one of the design scenarios, without adjustment the naïve MLE yielded a mean power of 78.8%, and by applying f from the Method 1 it increased to 82.5%, and to 86.2% and 87.5% by applying ρ from the Methods 1 and 2, respectively. Adjustment using ρ show results that are better and more consistent across different estimators and design scenarios as compared to using f. The ρ from the Method 2 attains mean power that is the closest to the target (90%). As in the previous simulations, the median power was higher than the mean power in all the cases.

Table 2 Mean and median power of Phase III trials planned using Phase II design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 20)$. The target power of the Phase III trial is of 90%.

		$f = 1$		$f = 0.9$		$f = 0.85$	
π	$\hat{\pi}$	Mean	Median	Mean	Median	Mean	Median
π_1	$\hat{\pi}_{nml}$	0.7884	0.8623	0.9105	0.9912	0.9495	0.9998
	$\hat{\pi}_{m1}$	0.8232	0.9050	0.9354	0.9971	0.9670	1.0000
	$\hat{\pi}_{m2}$	0.8250	0.9003	0.9356	0.9967	0.9670	1.0000
	$\hat{\pi}_{m2v2}$	0.8063	0.8752	0.9283	0.9934	0.9649	0.9999
	$\hat{\pi}_{m3}$	0.8248	0.8982	0.9356	0.9965	0.9670	1.0000
$\pi_1 + 0.1$	$\hat{\pi}_{nml}$	0.8075	0.8346	0.9341	0.9731	0.9699	0.9957
	$\hat{\pi}_{m1}$	0.8655	0.9005	0.9599	0.9913	0.9827	0.9993
	$\hat{\pi}_{m2}$	0.8656	0.9013	0.9599	0.9915	0.9827	0.9993
	$\hat{\pi}_{m2v2}$	0.8676	0.9013	0.9599	0.9915	0.9827	0.9993
	$\hat{\pi}_{m3}$	0.8656	0.9002	0.9599	0.9913	0.9827	0.9993

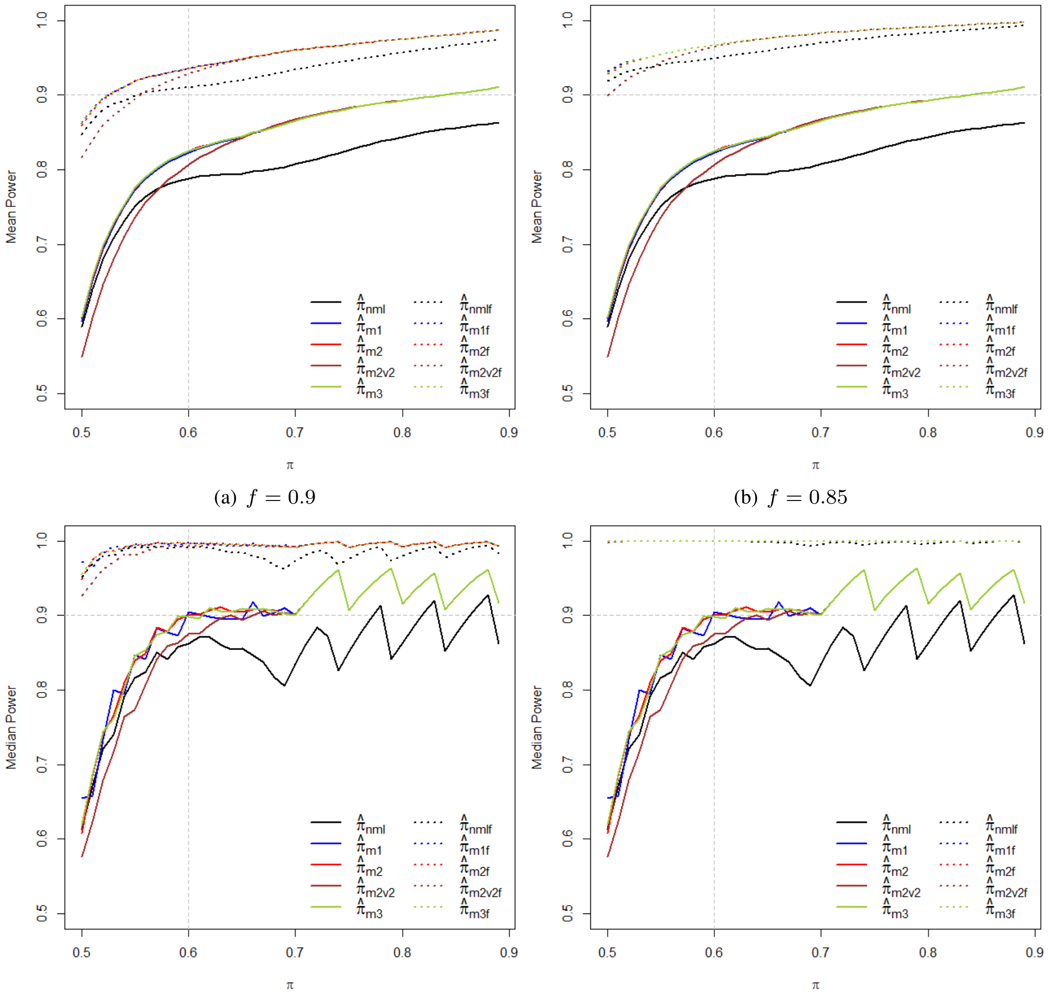
7 Discussion

In this paper we have studied the consequences of planning the Phase III sample size using the estimates from positive Phase II trials, taking oncology trials as a special case. In addition, we have proposed new methods to estimate adjustment factors based on the observed data.

The use of Phase II efficacy estimators that are less biased lead to a better Phase III power. However, as far as the average power is concerned, adjustments are necessary in order to reach the target value of the power irrespective of the efficacy estimator. These adjustments translate into retention (*discounting*) of the Phase II efficacy estimate. The extent of retention is dependant on the estimator, with better performing estimators requiring less reduction.

Although not reaching the target (nominal) average power, our proposed adjustment methods show improved results that are consistently similar for different estimators and design scenarios. Here we would recommend making adjustments using the Method 2, since simulations showed that it yields the best result, with the attained mean power being less than the target only by 3%. The consistence in results seen in this method suggests that it may also perform well in design types and estimators other than those considered in this paper.

We have assumed throughout this paper that Phase III sample size is estimated based solely on one Phase II treatment effect estimate. This is likely to be the case if the Phase II trial is the only source of information regarding treatment effect. In such a case, the focus of the investigators would be on how to adjust this single estimate in order to properly power the Phase III trial. In cases where multiple sources of information are available, other approaches might be more appropriate. One example is the meta-analytic approach proposed by Burke et al. (2014) for cases where data from various similar Phase II trials are available.



(a) $f = 0.9$

(b) $f = 0.85$

(c) $f = 0.9$

(d) $f = 0.85$

Figure 2 Mean and median power of Phase III trials planned using Phase II design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 20)$. The solid lines are for the unadjusted effect estimates, and the dashed lines for the adjusted ones (i.e., estimates multiplied by the retention factor f). For each value of π 50000 trials were simulated. The vertical line represents $\pi = \pi_1$ and the horizontal one represents the target power (90%).

Table 3 Multiplicative adjustment factors for effect size and sample size (f and ρ), and the corresponding attained power in Phase III trial. The first group rows correspond to the Phase II design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.2, 0.4, 0.05, 0.1, 20)$ and the other group to the design defined by $(\pi_0, \pi_1, \alpha, \beta, n_1) = (0.4, 0.6, 0.05, 0.1, 20)$.

(a) Adjustment factors

	Method 1				Method 2	
	f		ρ		ρ	
Est.	Mean	SD	Mean	SD	Mean	SD
$\hat{\pi}_{nml}$	0.951	0.034	1.591	0.367	1.731	0.393
$\hat{\pi}_{m1}$	0.982	0.047	1.406	0.348	1.513	0.356
$\hat{\pi}_{m2}$	0.983	0.048	1.414	0.371	1.526	0.385
$\hat{\pi}_{m2v2}$	0.980	0.057	1.427	0.382	1.548	0.409
$\hat{\pi}_{m3}$	0.983	0.047	1.413	0.369	1.524	0.382
$\hat{\pi}_{nml}$	0.966	0.017	1.591	0.320	1.706	0.308
$\hat{\pi}_{m1}$	0.984	0.026	1.405	0.320	1.492	0.311
$\hat{\pi}_{m2}$	0.984	0.026	1.411	0.339	1.502	0.333
$\hat{\pi}_{m2v2}$	0.982	0.034	1.450	0.394	1.572	0.416
$\hat{\pi}_{m3}$	0.985	0.026	1.409	0.336	1.499	0.330

(b) Power (target : 90%)

	No adjustment		Method 1				Method 2	
	pwr$[n(\hat{\theta}), \theta, \alpha]$		pwr$[n(f\hat{\theta}), \theta, \alpha]$		pwr$[\rho n(\hat{\theta}), \theta, \alpha]$		pwr$[\rho n(\hat{\theta}), \theta, \alpha]$	
Est.	Mean	Median	Mean	Median	Mean	Median	Mean	Median
$\hat{\pi}_{nml}$	77.8%	84.1%	80.7%	88.7%	85.2%	95.0%	86.6%	96.6%
$\hat{\pi}_{m1}$	82.1%	90.6%	81.4%	91.6%	85.9%	96.3%	87.3%	97.4%
$\hat{\pi}_{m2}$	82.3%	89.9%	81.5%	90.7%	85.9%	95.9%	87.3%	97.2%
$\hat{\pi}_{m2v2}$	81.3%	87.5%	80.4%	89.0%	85.1%	94.5%	86.7%	96.0%
$\hat{\pi}_{m3}$	82.3%	89.4%	81.4%	90.2%	85.9%	95.6%	87.3%	97.0%
$\hat{\pi}_{nml}$	78.8%	86.2%	82.5%	91.2%	86.2%	96.2%	87.5%	97.6%
$\hat{\pi}_{m1}$	82.2%	90.5%	82.6%	92.3%	86.3%	96.4%	87.5%	97.4%
$\hat{\pi}_{m2}$	82.5%	90.0%	82.7%	91.7%	86.4%	96.1%	87.6%	97.3%
$\hat{\pi}_{m2v2}$	80.3%	86.6%	80.3%	89.2%	84.4%	94.3%	86.0%	96.0%
$\hat{\pi}_{m3}$	82.4%	89.8%	82.7%	91.6%	86.4%	96.0%	87.6%	97.2%

Acknowledgements This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567.

Conflict of Interest *The authors have declared no conflict of interest.*

References

Ahn, C., Heo, M., & Zhang, S. (2014). *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. CRC Press.

Burke, D. L., Billingham, L. J., Girling, A. J., & Riley, R. D. (2014). Meta-analysis of randomized phase II trials to inform subsequent phase III decisions. *Trials*, *15*(1), 346.

Chuang-Stein, C. & Kirby, S. (2017). *Quantitative Decisions in Drug Development*. Springer.

De Martini, D. (2011a). Adapting by calibration the sample size of a phase III trial on the basis of phase II data. *Pharmaceutical Statistics*, *10*, 89–95.

De Martini, D. (2011b). Robustness and corrections for sample size adaptation strategies based on effect size estimation. *Communications in Statistics-Simulation and Computation*, *40*(9), 1263–1277.

De Martini, D. (2013). *Success probability estimation with applications to clinical trials*. John Wiley & Sons.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., & Verweij, J. (2009). New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer (Oxford, England : 1990)*, *45*, 228–247.

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

Gan, H. K., You, B., Pond, G. R., & Chen, E. X. (2012). Assumptions of expected benefits in randomized phase III trials evaluating systemic treatments for cancer. *Journal of the National Cancer Institute*, *104*, 590–598.

Halabi, S. (2008). Statistical considerations for the design and analysis of phase III clinical trials in prostate cancer. *Urologic Oncology: Seminars and Original Investigations*, *26*(3), 300 – 307. A Clinician's Guide to Statistical Methods in Urologic Oncology.

Kirby, S., Burke, J., Chuang-Stein, C., & Sin, C. (2012). Discounting phase 2 results when planning phase 3 clinical trials. *Pharmaceutical Statistics*, *11*, 373–385.

Nhacolo, A. & Brannath, W. (2018). Interval and point estimation in adaptive Phase II trials with binary endpoint. *Statistical Methods in Medical Research*, 096228021878141.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Saad, E. D., Katz, A., & Buyse, M. (2010). Overall survival and post-progression survival in advanced breast cancer: a review of recent randomized clinical trials. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, *28*, 1958–1962.

Shan, G., Wilding, G. E., Hutson, A. D., & Gerstenberger, S. (2016). Optimal adaptive two-stage designs for early phase II clinical trials. *Statistics in Medicine*, *35*(8), 1257–1266. sim.6794.

Wang, S.-J., Hung, H. M. J., & O'Neill, R. T. (2006). Adapting the sample size planning of a phase III trial based on phase II data. *Pharmaceutical Statistics*, *5*, 85–97.

Bibliography

A'Hern, R. P. (2004). Widening eligibility to phase II trials: constant arcsine difference phase II trials. *Controlled Clinical Trials*, 25(3):251–264.

Ahn, C., Heo, M., and Zhang, S. (2014). *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research.* CRC Press.

An, M.-W., Mandrekar, S. J., and Sargent, D. J. (2012). A 2-stage phase II design with direct assignment option in stage II for initial marker validation. *Clinical Cancer Research*, 18(16):4225–4233.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika*, 44(1/2):9–26.

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

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

Barker, A. D., Sigman, C. C., Kelloff, G. J., Hylton, N. M., Berry, D. A., and Esserman, L. J. (2009). I-spy 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*, 86:97–100.

Barroilhet, L. and Matulonis, U. (2018). The nci-match trial and precision medicine in gynecologic cancers. *Gynecologic Oncology.*

Bauer, P. and Brannath, W. (2004). The advantages and disadvantages of adaptive designs for clinical trials. *Drug Discovery Today*, 9(8):351–357.

Bebu, I., Dragalin, V., and Luta, G. (2013). Confidence intervals for confirmatory adaptive two-stage designs with treatment selection. *Biometrical Journal*, 55(3):294–309.

Bebu, I., Luta, G., and Dragalin, V. (2010). Likelihood inference for a two-stage design with treatment selection. *Biometrical Journal*, 52(6):811–822.

Berry, D. A. (2015). The brave new world of clinical cancer research: Adaptive biomarker-driven trials integrating clinical practice with clinical research. *Molecular oncology*, 9:951–959.

Bersimis, S., Sachlasb, A., and Papaioannou, T. (2015). Flexible designs for phase II comparative clinical trials involving two response variables. *Statistics in Medicine*, 34(2):197–214.

Biswas, A., Datta, S., Fine, J. P., and Segal, M. R., editors (2008). *Statistical Advances in the Biomedical Sciences – Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Bowden, J., Brannath, W., and Glimm, E. (2014). Empirical bayes estimation of the selected treatment mean for two-stage drop-the-loser trials: a meta-analytic approach. *Statistics in Medicine*, 33(3):388–400.

Bowden, J. and Glimm, E. (2008). Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal*, 50(4):515–527.

Bowden, J. and Glimm, E. (2014). Conditionally unbiased and near unbiased estimation of the selected treatment mean for multistage drop-the-losers trials. *Biometrical Journal*, 56(2):332–349.

Bowden, J. and Trippa, L. (2015). Unbiased estimation for response adaptive clinical trials. *Statistical Methods in Medical Research*.

Bowden, J. and Wason, J. (2012). Identifying combined design and analysis procedures in two-stage trials with a binary end point. *Statistics in Medicine*, 31(29):3874–3884.

Brannath, W., König, F., and Bauer, P. (2006). Estimation in flexible two stage designs. *Statistics in Medicine*, 25(19):3366–3381.

Broberg, P. and Miller, F. (2017). Conditional estimation in two-stage adaptive designs. *Biometrics*.

Brunier, H. C. and Whitehead, J. (1994). Sample sizes for phase II clinical trials derived from Bayesian decision theory. *Statistics in Medicine*, 13:2493–2502.

Brutti, P., Gubbiottib, S., and Sambucini, V. (2011). An extension of the single threshold design for monitoring efficacy and safety in phase II clinical trials. *Statistics in Medicine*, 30(14):1648–1664.

Bryant, J. and Day, R. (1995). Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*, 51(4):1372–1383.

Burke, D. L., Billingham, L. J., Girling, A. J., and Riley, R. D. (2014). Meta-analysis of randomized phase II trials to inform subsequent phase III decisions. *Trials*, 15(1):346.

Cai, C., Liu, S., and Yuan, Y. (2014). A bayesian design for phase II clinical trials with delayed responses based on multiple imputation. *Statistics in Medicine*, 33(23):4017–4028.

Carreras, M. and Brannath, W. (2013). Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statistics in Medicine*, 32(10):1677–1690.

Carsten, C. and Chen, P. (2015). Curtailed two-stage matched pairs design in double-arm phase II clinical trials. *Journal of Biopharmaceutical Statistics*.

Carter, G. M. and Rolph, J. E. (1974). Empirical bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, 69(348):880–885.

Case, L. D. and Morgan, T. M. (2003). Design of phase II cancer trials evaluating survival probabilities. *BMC Medical Research Methodology*, 3:6.

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

Chang, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics*, 45:247–254.

Chang, M. N. and O'Brien, P. C. (1986). Confidence intervals following group sequential tests. *Controlled Clinical Trials*, 7(1):18–26.

Chang, M. N., Shuster, J. J., and Kepner, J. L. (1999). Group sequential designs for phase II trials with historical controls. *Controlled Clinical Trials*, 20(4):353–364.

Chang, M. N., Therneau, T. M., Wieand, H. S., and Cha, S. S. (1987). Designs for group sequential phase II clinical trials. *Biometrics*, 43(4):865–874.

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

Chen, C.-M. and Chi, Y. (2012). Curtailed two-stage designs with two dependent binary endpoints. *Pharmaceutical Statistics*, 11(1):57–62.

Chen, K. and Shan, M. (2008). Optimal and minimax three-stage designs for phase II oncology clinical trials. *Contemporary Clinical Trials*, 29(1):32–41.

Chen, N. and Lee, J. J. (2013). Optimal continuous-monitoring design of single-arm phase II trial based on the simulated annealing method. *Contemporary Clinical Trials*, 35(1):170–178.

Chen, S., Soong, S., and Wheeler, R. H. (1994). An efficient multiple-stage procedure for phase II clinical trials that have high response rate objectives. *Controlled Clinical Trials*, 15(4):277–283.

Chen, T. T. (1997). Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*, 16(23):2701–2711.

Chen, T. T. and Ng, T.-H. (1998). Optimal flexible designs in phase II clinical trials. *Statistics in Medicine*, 17(20):2301–2312.

Chen, Y. and Smith, B. J. (2009). Adaptive group sequential design for phase II clinical trials: a Bayesian decision theoretic approach. *Statistics in Medicine*, 28(27):3347–3362.

Cheng, Y. and Shen, Y. (2004). Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics*, 60(4):910–918.

Cheung, Y. K. (2009). Selecting promising treatments in randomized phase II cancer trials with an active control. *Journal of Biopharmaceutical Statistics*, 19(3):494–508.

Chi, Y. and Chen, C.-M. (2008). Curtailed two-stage designs in phase II clinical trials. *Statistics in Medicine*, 27(29):6175–6189.

Chow, S.-C. and Chang, M. (2008). Adaptive design methods in clinical trials - a review. *Orphanet journal of rare diseases*, 3:11.

Chow, S.-C., Chang, M., and Pong, A. (2005). Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics*, 15:575–591.

Chuang-Stein, C. and Kirby, S. (2017). *Quantitative Decisions in Drug Development*. Springer.

Coburger, S. and Wassmer, G. (2003). Sample size reassessment in adaptive clinical trials using a bias corrected estimate. *Biometrical Journal*, 45(7):812825.

Cohen, A. and Sackrowitz, H. B. (1989). Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters*, 8(3):273–278.

Conaway, M. R. and Petroni, G. R. (1995). Bivariate sequential designs for phase II trials. *Biometrics*, 51(2):656–664.

Conaway, M. R. and Petroni, G. R. (1996). Designs for phase II trials allowing for a trade-off between response and toxicity. *Biometrics*, 52(4):1375–1386.

Cronin, K. A., Freedman, L. S., Lieberman, R., Weiss, H. L., Beenken, S. W., and Kelloff, G. J. (1999). Bayesian monitoring of phase II trials in cancer chemoprevention. *Journal of Clinical Epidemiology*, 52(8):705–711.

Cunanan, K. M., Iasonos, A., Shen, R., Begg, C. B., and Gönen, M. (2017). An efficient basket trial design. *Statistics in Medicine*, 36:1568–1579.

De Martini, D. (2011a). Adapting by calibration the sample size of a phase III trial on the basis of phase II data. *Pharmaceutical Statistics*, 10:89–95.

De Martini, D. (2011b). Robustness and corrections for sample size adaptation strategies based on effect size estimation. *Communications in Statistics-Simulation and Computation*, 40(9):1263–1277.

De Martini, D. (2013). *Success probability estimation with applications to clinical trials.* John Wiley & Sons.

Ding, M., Rosner, G. L., and Müller, P. (2008). Bayesian optimal design for phase II screening trials. *Biometrics*, 64(3):886–894.

Dong, G., Shih, W. J., Moore, D., Quan, H., and Marcella, S. (2012). A Bayesian-frequentist two-stage single-arm phase II clinical trial design. *Statistics in Medicine*, 31(19):2055–2067.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., and Verweij, J. (2009). New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer (Oxford, England : 1990)*, 45:228–247.

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77(4):875–892.

Englert, S. and Kieser, M. (2012a). Adaptive designs for single-arm phase II trials in oncology. *Pharmaceutical Statistics*, 11(3):241–249.

Englert, S. and Kieser, M. (2012b). Improving the flexibility and efficiency of phase II designs for oncology trials. *Biometrics*, 68(3):886–892.

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

Ensign, L. G., Gehan, E. A., Kamen, D. S., and Thall, P. F. (1994). An optimal three-stage design for phase II clinical trials. *Statistics in Medicine*, 13(17):1727–1736.

Estey, E. H. and Thall, P. F. (2003). New designs for phase 2 clinical trials. *Blood*, 102(2):442–448.

Fairbanks, K. and Madsen, R. (1982). P values for test using a repeated significance test design. *Biometrika*, 69(1):69–74.

Fan, X. F., Assaid, C. A., Ge, Y. J., and Ho, T. W. H. (2011). A two-stage adaptive design in phase 2 clinical trials for acute treatment of migraine. *Drug Information Journal*, 45(3):315–330.

Fan, X. F., DeMets, D. L., and Lan, K. K. G. (2004). Conditional bias of point estimates following a group sequential test. *Journal of Biopharmaceutical Statistics*, 14(2):505–530.

Ferrarotto, R., Redman, M. W., Gandara, D. R., Herbst, R. S., and Papadimitrakopoulou, V. A. (2015). Lung-map–framework, overview, and design principles. *Chinese clinical oncology*, 4:36.

Fleming, T. R. (1982). One-sample multiple testing procedure for phase II clinical trials. *Biometrics*, 38(1):143–151.

Gan, H. K., You, B., Pond, G. R., and Chen, E. X. (2012). Assumptions of expected benefits in randomized phase III trials evaluating systemic treatments for cancer. *Journal of the National Cancer Institute*, 104:590–598.

Gao, P., Liu, L., and Mehta, C. (2013). Exact inference for adaptive group sequential designs. *Statistics in Medicine*, 32(23):3991–4005.

Gehan, E. A. (1961). The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases*, 13(4):346–353.

Green, S. J. and Dahlberg, S. (1992). Planned versus attained design in phase II clinical trials. *Statistics in Medicine*, 11(7):853–862.

Guo, H. Y. and Liu, A. (2005). 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.

Halabi, S. (2008). Statistical considerations for the design and analysis of phase III clinical trials in prostate cancer. *Urologic Oncology: Seminars and Original Investigations*, 26(3):300 – 307. A Clinician's Guide to Statistical Methods in Urologic Oncology.

Hanfelt, J. J., Slack, R. S., and Gehan, E. A. (1999). 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.

Hee, S. W. and Stallard, N. (2012). Designing a series of decision-theoretic phase II trials in a small population. *Statistics in Medicine*, 31(30):4337–4351.

Heinrich, M. C., Joensuu, H., Demetri, G. D., Corless, C. L., Apperley, J., Fletcher, J. A., Soulieres, D., Dirnhofer, S., Harlow, A., Town, A., McKinley, A., Supple, S. G., Seymour, J., Di Scala, L., van Oosterom, A., Herrmann, R., Nikolova, Z., and McArthur, G. (2008). Phase II, open-label study evaluating the activity of imatinib in treating life-threatening malignancies known to be associated with imatinib-sensitive tyrosine kinases. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14:2717–2725.

Heitjan, D. F. (1997). Bayesian interim analysis of phase II cancer clinical trials. *Statistics in Medicine*, 16(16):1791–1802.

Herndon, J. E. I. (1998). A design alternative for two-stage, phase II, multicenter cancer clinical trials. *Controlled Clinical Trials*, 19(5):440–450.

Herson, J. (1979). Predictive probability early termination plans for phase II clinical trials. *Biometrics*, 35(4):775–783.

Herson, J. and Carter, S. K. (1986). Calibrated phase II clinical trials in oncology. *Statistics in Medicine*, 5:444–447.

Hilden, J. (1990). Corrected loss calculation for phase II trials. *Biometrics*, 46(2):535–538.

Hobbs, B. P., Chen, N., and Lee, J. J. (2018). Controlled multi-arm platform design using predictive probability. *Statistical Methods in Medical Research*, 27:65–78.

Hong, S. and Wang, Y. (2007). A three-outcome design for randomized comparative phase II clinical trials. *Statistics in Medicine*, 26(19):3525–3534.

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

Huang, X., Ning, J., Li, Y., Estey, E., Issa, J.-P., and Berry, D. A. (2009). Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Statistics in Medicine*, 28(12):1680–1689.

Hyman, D. M., Puzanov, I., Subbiah, V., Faris, J. E., Chau, I., Blay, J.-Y., Wolf, J., Raje, N. S., Diamond, E. L., Hollebecque, A., Gervais, R., Elez-Fernandez, M. E., Italiano, A., Hofheinz, R.-D., Hidalgo, M., Chan, E., Schuler, M., Lasserre, S. F., Makrutzki, M., Sirzen, F., Veronese, M. L., Tabernero, J., and Baselga, J. (2015). Vemurafenib in multiple nonmelanoma cancers with braf v600 mutations. *The New England journal of medicine*, 373:726–736.

Jennison, C. and Turnbull, B. W. (1983). Confidence intervals for a binomial parameter following a multistage test with application to mil-std 105d and medical trials. *Technometrics*, 15(1):49–58.

Jennison, C. and Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Chapman-Hall/CRC, Boca Raton, FL.

Jin, H. (2007). Alternative designs of phase II trials considering response and toxicity. *Contemporary Clinical Trials*, 28(4):525–531.

Jin, H. and Wei, Z. (2012). A new adaptive design based on Simon's two-stage optimal design for phase II clinical trials. *Contemporary Clinical Trials*, 33(6):1255–1260.

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

Jones, C. L. and Holmgren, E. (2007). An adaptive Simon two-stage design for phase 2 studies of targeted therapies. *Contemporary Clinical Trials*, 28(5):654–661.

Jovic, G. and Whitehead, J. (2010). An exact method for analysis following a two-stage phase II cancer clinical trial. *Statistics in Medicine*, 29(30):3118–3125.

Jung, S.-H. (2008). Randomized phase II trials with a prospective control. *Statistics in Medicine*, 27(4):568–583.

Jung, S.-H. (2013). *Randomized Phase II Cancer Clinical Trials*. Chapman & Hall/CRC Biostatistics Series. CRC Press, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742.

Jung, S.-H., Carey, M., and Kim, K. M. (2001). Graphical search for two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 22(4):367–672.

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

Jung, S.-H., Lee, T., Kim, K., and George, S. L. (2004). Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine*, 23(4):561–569.

Jung, S.-H. and Sargent, D. J. (2014). Randomized phase II clinical trials. *Journal of Biopharmaceutical Statistics*, 24(4):802–816.

Kaplan, R. (2015). The focus4 design for biomarker stratified trials. *Chinese clinical oncology*, 4:35.

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J., Gupta, S., Alden, C. M., Liu, S., Tang, X., Khuri, F. R., Tran, H. T., Johnson, B. E., Heymach, J. V., Mao, L., Fossella, F., Kies, M. S., Papadimitrakopoulou, V., Davis, S. E., Lippman, S. M., and Hong, W. K. (2011). The battle trial: personalizing therapy for lung cancer. *Cancer discovery*, 1:44–53.

Kimani, P. K., Todd, S., and Stallard, N. (2015). Estimation after subpopulation selection in adaptive seamless trials. *Statistics in Medicine*, 34(18):2581–2601.

Kirby, S., Burke, J., Chuang-Stein, C., and Sin, C. (2012). Discounting phase 2 results when planning phase 3 clinical trials. *Pharmaceutical Statistics*, 11:373–385.

Kocherginsky, M., Cohen, E. E. W., and Karrison, T. (2009). Design of phase II cancer trials for evaluation of cytostatic/cytotoxic agents. *Journal of Biopharmaceutical Statistics*, 19(3):524–529.

Koopmeiners, J. S., Feng, Z., and Pepe, M. S. (2012). Conditional estimation after a two-stage diagnostic biomarker study that allows early termination for futility. *Statistics in Medicine*, 31(5):420–435.

Koyama, T. and Chen, H. (2008). Proper inference from simon's two-stage designs. *Statistics in Medicine*, 27(16):3145–3154.

Kunz, C. U. and Kieser, M. (2011). Optimal two-stage designs for single-arm phase II oncology trials with two binary endpoints. *Methods of Information in Medicine*, 50(4):372–377.

Kunzmann, K. and Kieser, M. (2017). Point estimation and p-values in phase II adaptive two-stage designs with a binary endpoint. *Statistics in Medicine*, 36(6):971–984. sim.7200.

Kwak, M. and Jung, S.-H. (2014). Phase II clinical trials with time-to-event endpoints: optimal two-stage designs with one-sample log-rank test. *Statistics in Medicine*, 33(12):2004–2016.

Lai, X. and Zee, B. C.-Y. (2015). Mixed response and time-to-event endpoints for multistage single-arm phase II design. *Trials*, 16:250.

Leblanc, M., Rankin, C., and Crowley, J. (2009). Multiple histology phase II trials. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 15:4256–4262.

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

Lee, Y., Staquet, M., Simon, R., Catane, R., and Muggia, F. (1979). Two-stage plans for patient accrual in phase II cancer clinical trials. *Cancer Treatment Reports*, 63(11-12):1721–1726.

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

Li, Q. (2011). An mse-reduced estimator for the response proportion in a two-stage clinical trial. *Pharmaceutical Statistics*, 10(3):277–279.

Lin, J. and Bunn, V. (2017). Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials. *Contemporary Clinical Trials*, 54:48–59.

Lin, S. P. and Chen, T. T. (2000). Optimal two-stage designs for phase II clinical trials with differentiation of complete and partial responses. *Communications in Statistics-Theory and Methods*, 29(5-6):923–940.

Lin, X., Allred, R., and Andrews, G. (2008). A two-stage phase II trial design utilizing both primary and secondary endpoints. *Pharmaceutical Statistics*, 7(2):88–92.

Lin, Y. and Shih, W. J. (2004). Adaptive two-stage designs for single-arm phase IIA cancer clinical trials. *Biometrics*, 60(2):482–490.

Liu, A., Hall, W. J., Yu, K. F., and Wu, C. (2006). Estimation following a group sequential test for distributions in the one-parameter exponential family. *Statistica Sinica*, 16:165–181.

Liu, R., Liu, Z., Ghadessi, M., and Vonk, R. (2017). Increasing the efficiency of oncology basket trials using a bayesian approach. *Contemporary Clinical Trials*.

Logan, B. R. (2005). Optimal two-stage randomized phase II clinical trials. *Clinical Trials*, 2(1):5–12.

London, W. B. and Chang, M. N. (2005). One- and two-stage designs for stratified phase II clinical trials. *Statistics in Medicine*, 24(17):2597–2611.

Lu, Y., Jin, H., and Lamborn, K. R. (2005). A design of phase II cancer trials using total and complete response endpoints. *Statistics in Medicine*, 24(20):3155–3170.

Luo, X., Li, M., Shih, W. J., and Ouyang, P. (2012). Estimation of treatment effect following a clinical trial with adaptive design. *Journal of Biopharmaceutical Statistics*, 22(4):700–718.

Luo, X., Wu, S. S., and Xiong, J. (2010). Parameter estimation following an adaptive treatment selection trial design. *Biometrical Journal*, 52(6):823–835.

Magnusson, B. P. and Turnbull, B. W. (2013). Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine*, 32:2695–2714.

Mahajan, R. and Gupta, K. (2010). Adaptive design clinical trials: Methodology, challenges and prospect. *Indian journal of pharmacology*, 42:201–207.

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

Mander, A. P., M.S.Wason, J., Sweeting, M. J., and Thompson, S. G. (2012). 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.

Mariani, L. and Marubini, E. (1996). Design and analysis of phase II cancer trials: A review of statistical methods and guidelines for medical researchers. *International Statistial Review*, 64(1):61–88.

Masaki, N., Koyama, T., Yoshimura, I., and Hamada, C. (2009). Optimal two-stage designs allowing flexibility in number of subjects for phase II clinical trials. *Journal of Biopharmaceutical Statistics*, 19(4):721–731.

Mayo, M. S., Mahnken, J. D., and Soong, S.-J. (2010). Optimal designs for two-arm, phase II clinical trial design with multiple constraints. *Journal of Biopharmaceutical Statistics*, 20(1):106–124.

Milanzi, E., Molenberghs, G., Alonso, A., Kenward, M. G., Tsiatis, A. A., Davidian, M., and Verbeke, G. (2015). Estimation after a group sequential trial. *Stat Biosci*, 7(2):187–205.

Milanzi, E., Molenberghs, G., Alonso Abad, A., Kenward, M. G., Verbeke, G., Tsiatis, A. A., and Davidian, M. (2014). Properties of estimators in exponential family settings with observation-based stopping rules.

Molenberghs, G., Kenward, M. G., Aerts, M., Verbeke, G., Tsiatis, A. A., Davidian, M., and Rizopoulos, D. (2014). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research*, 23(1):11–41.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57:886–891.

Müller, H.-H. and Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine*, 23:2497–2508.

Neuenschwander, B., Wandel, S., Roychoudhury, S., and Bailey, S. (2016). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics*, 15:123–134.

Nhacolo, A. and Brannath, W. (2018). Interval and point estimation in adaptive Phase II trials with binary endpoint. *Statistical Methods in Medical Research*, page 096228021878141.

Panageas, K. S., Smith, A., Gönen, M., and Chapman, P. B. (2002). An optimal two-stage phase II design utilizing complete and partial response information separately. *Controlled Clinical Trials*, 23(4):367–379.

Pepe, M. S., Feng, Z., Longton, G., and Koopmeiners, J. (2009). Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility. *Statistics in Medicine*, 28(5):762–779.

Pinheiro, J. and DeMets, D. L. (1997). Estimating and reducing bias in group sequential designs with gaussian independent increment structure. *Biometrika*, 84(4):831–845.

Porcher, R. and Desseaux, K. (2012). What inference for two-stage phase II trials? *BMC Medical Research Methodology*, 12:117.

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C., and Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine*, 24(24):3697–3714.

Poulopoulou, S., Karlis, D., Yiannoutsos, C. T., and Dafni, U. (2014). Phase II design with sequential testing of hypotheses within each stage. *Journal of Biopharmaceutical Statistics*, 24(4):768–784.

Pretorius, A. G. (2016). Phase III trial failures: Costly, but preventable. *Applied Clinical Trials*, 25(8).

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R., Miller, J. P., and Rao, D. C. (2007). *Epidemiology and medical statistics*, volume 27. Elsevier.

Ray, H. and Rai, S. (2011). An evaluation of a simon 2-stage phase II clinical trial design incorporating toxicity monitoring. *Contemporary Clinical Trials*, 32(3):428–436.

Ray, H. E. and Rai, S. N. (2012). Operating characteristics of a simon two-stage phase II clinical trial design incorporating continuous toxicity monitoring. *Pharmaceutical Statistics*, 11(2):170–176.

Ray, H. E. and Rai, S. N. (2013). Flexible bivariate phase II clinical trial design incorporating toxicity and response on different schedules. *Statistics in Medicine*, 32(3):470–485.

Renfro, L. A. and Sargent, D. J. (2017). Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. *Annals of oncology : official journal of the European Society for Medical Oncology*, 28:34–43.

Roberts, J. D. and Ramakrishnan, V. (2011). Phase II trials powered to detect tumor subtypes. *Clinical Cancer Research*, 17(17):5538–5545.

Rosner, G. L. and Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika*, 75(4):723–729.

Saad, E. D., Katz, A., and Buyse, M. (2010). Overall survival and post-progression survival in advanced breast cancer: a review of recent randomized clinical trials. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28:1958–1962.

Sambucini, V. (2008). A Bayesian predictive two-stage design for phase II clinical trials. *Statistics in Medicine*, 27(8):1199–1224.

Sambucini, V. (2010). A Bayesian predictive strategy for an adaptive two-stage design in phase II clinical trials. *Statistics in Medicine*, 29(13):1430–1442.

Sargent, D. J., Chan, V., and Goldberg, R. M. (2001). A three-outcome design for phase II clinical trials. *Controlled Clinical Trials*, 22(2):117–125.

Saville, B. R. and Berry, S. M. (2016). Efficiencies of platform clinical trials: A vision of the future. *Clinical trials (London, England)*, 13:358–366.

Schultz, J. R., Nichol, F. R., Elfring, G. L., and Weed, S. D. (1973). Multiple-stage procedures for drug screening. *Biometrics*, 29(2):293–300.

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

Shan, G., Wilding, G. E., Hutson, A. D., and Gerstenberger, S. (2016a). Optimal adaptive two-stage designs for early phase II clinical trials. *Statistics in Medicine*, 35(8):1257–1266. sim.6794.

Shan, G., Zhang, H., and Jiang, T. (2016b). Minimax and admissible adaptive two-stage designs in phase II clinical trials. *BMC Medical Research Methodology*, 16(1):90.

Shen, L. (2001). An improved method of evaluating drug effect in a multiple dose clinical trial. *Statistics in Medicine*, 20(13):19131929.

Shuster, J. (2002). Optimal two-stage designs for single arm phase II cancer trials. *Journal of Biopharmaceutical Statistics*, 12(1):39–51.

Siegmund, D. (1978). Estimation following sequential tests. *Biometrika*, 65(2):341–349.

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

Simon, R., Geyer, S., Subramanian, J., and Roychowdhury, S. (2016). The bayesian basket design for genomic variant-driven phase II trials. *Seminars in Oncology*, 43:13–18.

Song, J. X. (2014). A two-stage patient enrichment adaptive design in phase II oncology trials. *Contemporary Clinical Trials*, 37(1):148–154.

Song, J. X. (2015). A two-stage design with two co-primary endpoints. *Contemporary Clinical Trials*, 1:2–4.

Sridhara, R., He, K., Nie, L., Shen, Y.-L., and Tang, S. (2015). Current statistical challenges in oncology clinical trials in the era of targeted therapy. *Statistics in Biopharmaceutical Research*, 7(4):348–356.

Stallard, N. (1998). Sample size determination for phase II clinical trials based on bayesian decision theory. *Biometrics*, 54(1):279–294.

Stallard, N. (2003). Decision-theoretic designs for phase II clinical trials allowing for competing studies. *Biometrics*, 59(2):402–409.

Stallard, N. and Cockey, L. (2008). Two-stage designs for phase II cancer trials with ordinal responses. *Contemporary Clinical Trials*, 29(6):896–904.

Stallard, N., Thall, P. F., and Whitehead, J. (1999). Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics*, 55(3):971–977.

Stallard, N. and Todd, S. (2005). Point estimates and confidence regions for sequential trials involving selection. *Journal of Statistical Planning and Inference*, 135(2):402419.

Stallard, N., Todd, S., and Whitehead, J. (2008). Estimation following selection of the largest of two normal means. *Journal of Statistical Planning and Inference*, 138(6):16291638.

Storer, B. E. (1992). A class of phase II designs with three possible outcomes. *Biometrics*, 48(1):55–60.

Su, Z. (2010). An adaptive design for phase II non-oncology dose selection clinical trials. *Clinical Drug Investigation*, 30(6):397–403.

Sun, L. Z., Chen, C., and Patel, K. (2009). Optimal two-stage randomized multinomial designs for phase II oncology trials. *Journal of Biopharmaceutical Statistics*, 19(3):485–493.

Sylvester, R. J. (1988). A Bayesian approach to the design of phase II clinical trials. *Biometrics*, 44(3):823–836.

Tamura, R. N., Huang, X., and Boos, D. D. (2011). Estimation of treatment effect for the sequential parallel design. *Statistics in Medicine*, 30(30):3496–3506.

Tan, M. T. and Xiong, X. (2011). A flexible multi-stage design for phase II oncology trials. *Pharmaceutical Statistics*, 10(4):369–373.

Tan, S.-B. and Machin, D. (2002). Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine*, 21(14):1991–2012.

Tan, S.-B. and Machin, D. (2006). Letter to the editor: Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine*, 25(19):3407–3408.

Thall, P. F. and Simon, R. (1990). Incorporating historical control data in planning phase II clinical trials. *Statistics in Medicine*, 9:215–228.

Thall, P. F. and Simon, R. (1994). Practical bayesian guidelines for phase IIB clinical trials. *Biometrics*, 50(2):337–349.

Thall, P. F., Simon, R., and Ellenberg, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika*, 75(2):303–310.

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

Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H., and Benjamin, R. S. (2003). Hierarchical bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine*, 22:763–780.

Therneau, T. M., Wieand, H., and Chang, M. (1990). Optimal designs for a grouped sequential binomial trial. *Biometrics*, 46(3):771–781.

Todd, S., Whitehead, J., and Facey, K. M. (1996). Point and interval estimation following a sequential clinical trial. *Biometrika*, 83(2):453–461.

Tournoux-Facon, C., De Rycke, Y., and Tubert-Bitter, P. (2011). Targeting population entering phase III trials: a new stratified adaptive phase II design. *Statistics in Medicine*, 30(8):801–811.

Tsai, W.-Y., Chi, Y., and Chen, C.-M. (2008). Interval estimation of binomial proportion in clinical trials with a two-stage design. *Statistics in Medicine*, 27(1):15–35.

Tsiatis, A. A., Rosner, G. L., and Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics*, 40(3):797–803.

Wang, S.-J., Hung, H. M. J., and O'Neill, R. T. (2006). Adapting the sample size planning of a phase III trial based on phase II data. *Pharmaceutical Statistics*, 5:85–97.

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

Wason, J. M. S. and Mander, A. P. (2012). Minimizing the maximum expected sample size in two-stage phase II clinical trials with continuous outcomes. *Journal of Biopharmaceutical Statistics*, 22(4):836–852.

Wassmer, G. and Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer Series in Pharmaceutical Statistics. Springer International Publishing.

Whitehead, J. (1985). Designing phase II studies in the context of a programme of clinical research. *Biometrics*, 41(2):373–383.

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

Whitehead, J. (2014). One-stage and two-stage designs for phase II clinical trials with survival endpoints. *Statistics in Medicine*, 33(22):3830–3843.

Whitehead, J., Valdés-Márquez, E., and Lissmats, A. (2009). A simple two-stage design for quantitative responses with application to a study in diabetic neuropathic pain. *Pharmaceutical Statistics*, 8(2):125–135.

Wilding, G. E., Shan, G., and Hutson, A. D. (2012). Exact two-stage designs for phase II activity trials with rank-based endpoints. *Contemporary Clinical Trials*, 33(2):332–341.

Woodcock, J. and LaVange, L. M. (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine*, 377(1):62–70.

Wu, C. and Liu, A. (2007). An adaptive approach for bivariate phase II clinical trial designs. *Contemporary Clinical Trials*, 28(4):482–486.

Wunder, C., Kopp-Schneider, A., and Edler, L. (2012). An adaptive group sequential phase II design to compare treatments for survival endpoints in rare patient entities. *Journal of Biopharmaceutical Statistics*, 22(2):294–311.

Ye, F. and Shyr, Y. (2007). Balanced two-stage designs for phase II clinical trials. *Clinical Trials*, 4(5):514–524.

Yuan, S. S., Chen, A., He, L., Chen, C., Gause, C. K., and Beckman, R. A. (2016). On group sequential enrichment design for basket trial. *Statistics in Biopharmaceutical Research*, 8(3):293–306.

Zee, B., Melnychuk, D., Dancey, J., and Eisenhauer, E. (1999). Multinomial phase II cancer trials incorporating response and early progression. *Journal of Biopharmaceutical Statistics*, 9(2):351–363.

Zhang, Y., Mietlowski, W., Chen, B., and Wang, Y. (2011). An efficient algorithm to determine the optimal two-stage randomized multinomial designs in oncology clinical trials. *Journal of Biopharmaceutical Statistics*, 21(1):56–65.

Zhao, L., Taylor, J. M. G., and Schuetze, S. M. (2012). Bayesian decision theoretic two-stage design in phase II clinical trials with survival endpoint. *Statistics in Medicine*, 31(17):1804–1820.

Zhao, L. and Woodworth, G. (2009). Bayesian decision sequential analysis with survival endpoint in phase II clinical trials. *Statistics in Medicine*, 28(9):1339–1352.

Zhong, B. (2012). Single-arm phase IIA clinical trials with go/no-go decisions. *Contemporary Clinical Trials*, 33(6):1272–1279.

Zhong, W., Koopmeiners, J. S., and Carlin, B. P. (2013). A two-stage Bayesian design with sample size reestimation and subgroup analysis for phase II binary response trials. *Contemporary Clinical Trials*, 36(2):587–596.

Zhong, W. and Zhong, B. (2013). One-sample proportion testing procedures for hypothesis of inequality. *Journal of Biopharmaceutical Statistics*, 23(3):604–617.