Doctoral Thesis

PhD in Pure and Applied Mathematics (31^{st} Cycle)

Innovative adaptive designs in oncology clinical trials with drug combinations

José L. Jiménez

Thesis Supervisor:

Prof. Mauro Gasparini

Thesis defended on **November 30th, 2018**

Doctoral Examination Committee:

Prof. Elizabeth Garrett-Mayer (Medical University of South Carolina, U.S.A)	Referee
Prof. Nolan Wages (University of Virginia, U.S.A)	Referee
Prof. Ziad Taib (Chalmers University of Technology, Sweden)	Examiner
Prof. Federico Ambrogi (Università degli Studi di Milano, Italy)	Examiner
Prof. Jean-Marie Grouin (Université de Rouen, France)	Examiner
Prof. Mauro Gasparini (Politecnico di Torino, Italy)	Thesis Supervisor

Politecnico di Torino

2018

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

José L. Jiménez
2018

* This dissertation is presented in partial fulfillment for the degree of Doctor of Philosophy.

Me gustaría dedicar esta tesis a mi madre, a mi padre y a mi hermano.

Acknowledgements

First of all, I have to thank Mauro Gasparini for me giving me the opportunity of joining IDEAS, the freedom to pursue the research topics I found more interesting and for teaching me how to be an independent researcher. I also would like to thank all the IDEAS supervisors for the training and the valuable advice you gave me along these three years.

Probably one of the most difficult things for me to do is how to thank Mourad Tighiouart. You not only hosted me five months in Los Angeles but and spent a lot of hours discussing new ideas, but most important, you shared with me your knowledge and passion for what you do and inspired me to continue working in this field. From the very bottom of my heart, thank you.

I am also in debt with Marcio Diniz, André Rogatko and Silke Rogatko for welcoming at your homes during my last visit to Los Angeles. You made me feel at home.

I also want to thank Byron Jones and Vika Stalbovskaya for hosting me at Novartis. It was a pleasure to shadow you Vika. Thank you for all the discussions we had about how clinical trials are done in industry. You passion for what you do was really inspiring and helped me to decide what I wanted to do after the PhD.

I also want to thank Thomas Jaki and Franz König for your help in the development of one of my papers, but also for organizing such a wonderful group of people.

I am thankful to the IDEAS colleagues I had during these years. I had so much fun with you in all our trips together. I will miss spending time with you in those fancy hotels. I want to thank Pavel, Enya, Haiyan, Johanna and Marius for being such great hosts in my visits to Lancaster and Basel respectively. To Nico, in particular for telling me about IDEAS back in 2014; ha sido un placer haber sido tu compañero durante estos tres años. To Maritina, for all the fun we had together, specially in Paris. And to Fabi, for all the endless conversations, the support, and the fun we have had in these three years.

Special thanks to Gaëlle and Elvira for the support during these years in Turin. I will always treasure those rare moments where the three of us where together in Turin.

Quiero también dar las gracias al que fue mi tutor durante mis estudios de máster, José Dorronsoro. El trabajo que hicimos juntos durante dos años me ha permitido acabar este doctorado y disfrutar durante el transcurso del mismo.

Quiero dar las gracias a mi amigo Marco, a toda la gente que he conocido de la Sociedad Española de Biometría, en particular al grupo de Barcelona, y a Carlos, Celia, Dani y Ainoa. Os conozco desde hace 16, 9, 13 y 10 años respectivamente y siempre habéis estado a mi lado en los momentos fáciles y en los momentos difíciles. No sé qué haría sin vosotros y vosotras.

También quiero dar las gracias a mi tío Fernando y a mi tía Rosa por apoyarme tanto, y por supuesto, a mis abuelos Mariano y Margarita; sé que estaríais orgullosos de mi.

Me cuesta encontrar palabras de agradecimiento para mi madre, mi padre y mi hermano, a quien esta tesis va dedicada. Gracias por haberme ayudado a llegar hasta aquí.

Last, and perhaps for the last time, 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.

Abstract

The use of drug combinations in clinical trials has emerged during the last years as an alternative to single agent trials since a more favorable therapeutic response may be obtained by combining drugs that, for instance, target multiple pathways or inhibit resistance mechanisms. This practice is common in both early phase and late phase clinical trials. However, depending on the phase of the trial, we may find different challenges that will require novel methodology. In early phase, where we model the probability of toxicity and efficacy, the main challenge is to find a suitable multivariate model that works well with a relatively low sample size. In late phase trials, the main challenge is to propose a design that allows to perfectly control the the type-I error and the power while allowing for the trial to stop in case of a lack of efficacy or in case the interim analyses show an efficacy that is big enough so it would be unethical to continue the trial. Other challenges may involve certain characteristics of the drug, such us delayed effects. This issue is quite present in nowadays clinical research because of the use of immuno-therapy against cancer.

In early phase trials, we studied the state of the art methodology and we observed that a large number of published methods are not appropriate for drug combination settings since were originally designed for single agents and then adapted to drug combinations. This statement is not based only on performance, because in fact many of these methods perform quite well even though they were not designed to be used in a drug combination setting, but because most of them do not take into account the interaction between drugs.

In late phase trials we focused our attention on the design of clinical trials in the presence of delayed effects in a drug combination setting. We performed a state of the art methodology review, and we observed that there is enough published methodology to design efficient confirmatory trials under this conditions. However, we also observed that most of this methodology primarily focuses on power recovery rather than type-I error rate control, which makes it difficult to apply in practice given the nature of confirmatory trials.

Our intention during this thesis was not only to develop novel methodology but to do it in areas that could be of interest for clinicians. In this thesis we make three contributions to the

field of clinical trials with drug combinations. In early phase trials, we propose a Bayesian adaptive phase I trial design that allows the investigator to attribute a DLT to one or both agents in a unknown fraction of patients, even when the drugs are given concurrently. We also propose a Bayesian adaptive phase I/II design with drug combinations, a binary endpoint in stage 1, and a TTP endpoint in stage 2, where we aim to identify the dose combination region associated with the highest median TTP among doses along the MTD curve. In late phase trials, we did an assessment of the impact of delayed effects in group sequential and adaptive group sequential designs, with an empirical evaluation in terms of power and type-I error rate of the weighted log-rank in a simulated scenario. Our last contribution includes several practical recommendations regarding which methodology should be used in the presence of delayed effects depending on certain characteristics of the trial.

Abstract

L'uso di combinazioni di farmaci è aumentato considerevolmente durante gli ultimi anni come alternativa agli studi clinici con un singolo farmaco. Ciò è dovutto alla possibilità di ottenere una migliore risposta individuando molteplici pathway oppure inibendo i meccanismi di resistenza. Questa pratica è comune in tutte le fasi dello sviluppo clinico. Tuttavia, ogni fase clinica presenta diverse sfide che richiedono lo sviluppo di nuove tecniche metodologiche. Nella fase iniziale dello sviluppo clinico, dove l'obiettivo è modellare la probabilità di tossicità ed efficacia, la difficoltà principale è trovare un modello soddisfacente con un numero ridotto di pazienti. Nella fase avanzata dello sviluppo clinico, l'obiettivo è controllare l'errore di tipo-I e la potenza statistica, e avere la flessibilità di interrompere lo studio nel caso di mancanza di efficacia oppure nel caso di risultati intermedi con sufficiente evidenza. Altre sfide riguardano per esempio sono la presenza di effetti ritardati. Questa situazione è molto comune nella ricerca clinica per l'uso dell'immuno-terapia.

Nella fase iniziale dello sviluppo clinico abbiamo studiato la metodologia finora presente in letteratura e abbiamo osservato che molte pubblicazioni non sono appropriate per l'uso di combinazioni di farmaci, perchè originalmente sono pensate per studi clinici con un singolo farmaco, e poi adattate alle combinazioni di farmaci. Questa affermazione non è solo basata sulle performance, ma anche sul fatto che queste metodologie non incorporano l'interazione di farmaci.

Nella fase avanzata dello sviluppo clinico ci siamo focalizzati sul disegno di studi clinici con effetti ritardati. Abbiamo studiato la metodologia principale e abbiamo concluso che esiste un numero sufficiente di techniche per il disegno di studi clinici con effetti retardati. Tuttavia, tutte le metodologie presenti sono incentrate sul recupero della potenza statistica trascurando il controllo dell'errore di I tipo, cosa che ne rende difficile l'applicabilità, data la natura degli studi clinici di conferma.

In questa tesi di dottorato abbiamo sviluppato tre nuovi contributi in aree di interesse per la ricerca clinica. Abbiamo realizzato tre articoli, due nell'area della fase iniziale e uno nell'area della fase avanzata. Abbiamo proposto un disegno adattativo Bayesiano per la fase

I che permette l'attribuzione di tossicità a uno o più farmaci. Abbiamo anche proposto un disegno adattativo Bayesiano di fase I/II, con endpoint binario nella prima fase e endpoint di sopravvivenza nella seconda fase dove vogliamo trovare la regione con la mediana di tempo fino alla progressione (TTP) più alta. Nella fase avanzata, abbiamo fatto un studio sull'impatto degli effetti ritardati nei disegni adattivi e nei disegni sequenziali adattivi. Questa valutazione riguarda la potenza statistica e l'errore di I tipo utilizzando il test ponderato dei ranghi logaritmici. Quest'ultimo lavoro include molte raccomandazioni pratiche da usare nel disegno di studi clinici con effetti ritardati.

Scientific Production

Journal Articles

Published / Accepted / Under review articles

- **Jimenez JL**, Tighiouart M, Gasparini M. Cancer phase I trial design using drug combinations when a fraction of dose limiting toxicities is attributable to one or more agents. *Biometrical Journal*. 2018; 1-15.
- **Jimenez JL**, Kim S, Tighiouart M. A Bayesian two-stage adaptive design for cancer phase I/II trials with drug combinations. *Under Review in Pharmaceutical Statistics.*
- **Jimenez JL**, Stalbovskaya V, Jones B. Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects. *Accepted in Pharmaceutical Statistics.*

Ongoing work

- **Jimenez, JL.** and Tighiouart, M. Bayesian latent variable modeling in cancer phase I/II clinical trials with drug combinations.

International Conferences

- 3rd European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) Workshop on Regulatory Statistics, Basel, Switzerland, September 24 - 25, 2018.
- XXIXth International Biometric Conference (29th IBC), Barcelona, Spain, July 8-13, 2018.
- BAYES 2018: Bayesian Biostatistics Conference, Cambridge, United Kingdom, June 20 - 22, 2018.
- 38th Annual Conference of the International Society of Clinical Biostatistics (ISCB), Vigo, Spain, July 09-13, 2017.
- Statisticians in the Pharmaceutical Industry (PSI) Conference, London, United Kingdom, May 14 - 17, 2017.

International Collaborations

- Mourad Tighiouart, Cedars-Sinai Medical Center, Los Angeles (CA), USA.
- Byron Jones, Novartis Pharma A.G., Basel, Switzerland.
- Viktoriya Stalbovskaya, Novartis Pharma A.G., Basel, Switzerland.

Contents

Chapter 1

Introduction

1.1 Drug development

Drug development is a long and costly process. It may take several years since a molecule is discovered until it is approved for sale by the regulatory agencies. In general, this process can be divided in 5 steps:

1. Pre-clinical studies.
2. Phase I studies.
3. Phase II studies.
4. Phase III studies.
5. Phase VI studies.

The goal of pre-clinical studies consists on identifying and studying the structure of new molecules and the affect they have on cells. This process serves as screening and those molecules that present activity on cells and then administered to laboratory animals in order to obtain estimate of the effect.

Pre-clinical studies are particularly useful because they are dose relatively fast and with low risk, since they are done in animals. They are strongly controlled, given the laboratory environment in which the are performed, and provide a detailed description of the mechanism of action for the animal model. Moreover, they allow us to gather safety data in order to evaluate the tolerability of the molecule. Pharmacokinetics / pharmacodynamics (PK / PD)

data is also collected to study the behavior of the molecule after its administration in the body.

Very few molecules make it past this phase of research. The transition from pre-clinical studies to phase I studies, is not straightforward, even though we already have some safety and efficacy information from animal studies. However, despite that the collected information is just a hint of the effect that the molecule may have in humans, the first dose level that will be used in humans is determined from the observed pre-clinical data.

Phase I studies are the first stage of testing the molecule, or drug, in humans. The goal of phase I trials is to evaluate the safety of the drug and identify the maximum tolerated dose (MTD). For non-life-threatening diseases, phase I studies are usually conducted on healthy human volunteers. However, in life-threatening diseases such as cancer, these studies are conducted on actual patients. The reason is due to the aggressiveness and potential treatment side effects, but also due to a high interest in administering the drug to patients for which there is no other therapeutic alternative.

Once the initial safety of the study drug have been tested and an MTD set has been identified, phase II trials are performed. The purpose of these studies is to identify therapeutic areas in which the drug presents promising efficacy results and also confirm the safety results found in precious phase I trials. If the phase II is successful, a phase III trial is performed. These trials compare the efficacy of the drug with current "gold standard" treatment or a placebo, depending on the therapeutic area. They are usually randomized controlled multi-center trials on large patient groups and are highly regulated by the regulatory agencies such as the European Medicines Agency (EMA) or the U.S. Food and Drug Administration (FDA).

Last, phase IV trials involve the safety surveillance and ongoing support of a drug after it receives permission to be marketed. Their purpose consist of collect, detect and monitor adverse events (AE) in already marketed drugs. These trials allow to detect rare or late adverse events in the general population that were not identified in previous clinical trials.

1.2 Dose finding clinical trials with drug combinations

In this section we present a brief review of adaptive phase I and II clinical trials as necessary tools for dose finding in drug development. Even though some methodology for single agent trials may be described, the main interest lies on methodology for drug combination trials. Moreover, because this thesis is focused on adaptive methodology, only model based designs

where the dose - toxicity and the dose - efficacy relationships are modeled using a statistical model are reviewed.

In oncology dose finding clinical trials, the main goal is to identify a dose that maximizes the treatment efficacy. The traditional approach would be to plan two different clinical trials: first a phase I trial where the goal is to obtain an estimate of the MTD, and then a phase II trial where among the set of doses recommended from the phase I, we look for the dose with highest probability of efficacy. An alternative would be to put both phase I and II trials into a unique phase I/II trial where toxicity and efficacy could be jointly modeled. However, this is only possible in settings where efficacy is observed relatively fast (e.g. one or two cycles of therapy). In cases where efficacy is not ascertained in a short period of time, phase I/II trials can also be employed but it is frequent to make use of two-stage designs where, an MTD is selected in the first stage of the trial, and then tested for efficacy in the second stage with possibly a different population than the one used in the first stage.

1.2.1 Phase I

Cancer phase I clinical trials constitute the first step in investigating a potentially promising drug. Due to safety and ethical concerns, patients have to be sequentially enrolled in the trial, and the dose combinations assigned to subsequent patients depend on dose combinations already given to previous patients and their DLT status. The main objective of phase I trials is to estimate an MTD that will be used in further efficacy evaluation. The MTD is usually defined as any dose combination (x,y) that will produce DLT in a prespecified proportion θ of patients,

$$P(\text{DLT}|\text{dose} = (x,y)) = \theta. \tag{1.1}$$

The definition of DLT depends on the type of cancer and drugs under study, but it is usually defined as a grade 3 or 4 non-hematologic toxicity (see the National Cancer Institute CTCAE v4.03 for the definition of the different grades of toxicity). The pre-specified proportion of DLTs θ, sometimes referred as target probability of DLT, also depends on the nature of the toxicity.

Continuar reassessment method (CRM)

The continual reassessment method is considered one of the first model-based phase I designs in the history of clinical trials and it is particularly relevant in oncology trials given the consequences of an inappropriate given dose [1, Chapter 1]. It was developed by [2] as an alternative to the traditional ruled-based designs. As stated by [3, Chapter 6], the CRM addresses ineffectiveness of treatment at low doses, severe toxic effects expected at high doses, poor knowledge of the dose-toxocity relationship at the trial onset, or the need for efficient designs with small sample sizes.

CRM is considered the first Bayesian adaptive design for dose finding because it uses all the accumulated data, including the data prior to the trial onset, at the time each dose level is estimated for new patients.

In this chapter we review the CRM, including the one-stage and two-stage variations. These single-agent strategies will serve as the basis for the CRM when used for drug combination.

In its most basic form the CRM characterizes the dose-toxicity relationship by a one-parameter parametric model, such us the hyperbolic tangent model, the logistic model, or the power model among others. In a more general expression, let (x,y) be the administered dose combination. The probability of DLT is given by

$$P(\text{DLT}|x,y) = F(\beta_0 + \beta_1 x + \beta_2 y + \beta_3 xy), \tag{1.2}$$

where F is a known cumulative distribution function.

In a Bayesian framework, uncertainty over the β parameters is expressed through a prior probability density function. For instance, [2] proposed to use an exponential prior centered on 1 and hence, $P(\beta) = \exp^{-\beta}$, where $0 < \beta < \infty$.

Let T_j be the binary toxicity outcome for the j-*th* patient, where $T_j = 1$ if a DLT is observed, and $T_j = 0$ otherwise. The likelihood function is defined as

$$L(\beta|x,y,T) \propto \prod_{j=1}^{n} [P(\text{DLT}|x,y)]^{T_j} \times [1 - P(\text{DLT}|x,y)]^{1-T_j}, \tag{1.3}$$

and the posterior probability distribution of the model parameters as

$$P(\beta_0,\beta_1,\beta_2,\beta_3|x,y,T) \propto P(\beta_0,\beta_1,\beta_2,\beta_3) \times L(\beta|x,y,T). \tag{1.4}$$

With (1.4) we can easily sample and obtain MCMC samples of the β parameters. The dose combination that will be recommended to the next patient to enter the trial is calculated as

$$(x_{\text{new}}, y_{\text{new}}) = \underset{(x,y)}{argmin} \, |\hat{P}(\text{DLT}|x,y) - \theta|, \tag{1.5}$$

where θ represents our target probability. We repeat this step until we reach the maximum sample size. Using the definition of MTD given by equation (1.1) and the dose-toxicity model defined in equation 1.2, at the end of the trial we estimate the MTD as follows

$$\hat{C} = \left\{ (x,y) : y = \frac{F^{-1}(\theta) - \hat{\beta}_0 - \hat{\beta}_1 x}{\hat{\beta}_2 + \hat{\beta}_3 x} \right\} \tag{1.6}$$

Even though CRM obtains accurate estimates of β, and hence more patients are treated at dose levels that are close to the MTD, it was not well accepted in its original form, mainly due to safety considerations. For this reason, as stated in [4], the original CRM was modified in order to add safety measures such us treating the first patient at the lowest starting dose level based on animal toxicology and conventional criteria, increasing the dose by only pre-specified level at a time, not allowing dose escalation for the immediate next patient if a patient has experienced DLT, or treating more than one patient at the same dose level, specially at high dose levels.

This methodology has been extensively studied, for instance, by [5–10].

Escalation with overdose control (EWOC)

As mentioned by [4, 3], the CRM may expose patients to overly toxic doses if either the model is misspecified or the first patient responses are atypical. Escalation with overdose control (EWOC) represents the first statistical method that incorporates a safety constraint into the design of the clinical trial, allowing more patients to be treated with potentially therapeutic doses [11]. This method is very similar to CRM. However, while CRM always uses the median of the MTD's posterior distribution to recommend subsequent doses, EWOC recommends the dose that is at the α-th percentile of the MTD's posterior distribution. α is also knows as a feasibility bound and it usually takes values between 0.25 and 0.5.

EWOC has been extensively studied, for example, by [12–14].

1.2.2 Phase II

After an MTD has been selected from a phase I trial, we proceed to run a phase II trial to evaluate if the dose (or doses) selected MTD has enough activity and also we aim to increase our knowledge about the toxicity of the drug. It is important to distinguish between two types of phase II trials: phase IIA and phase IIB.

Phase IIA trials are the initial efficacy evaluation and are usually designed as sigle-arm multi-stage design. They usually have a binary efficacy endpoint, they treat between 40 and 100 patients and they have early stopping rules to stop the trial in case of an obvious lack of efficacy at interim analyses.

Phase IIB trials are the subsequent efficacy evaluation and are usually multi-arm randomized multi-stage trials with the aim of identifying the most promising dose among those selected as MTD. In this kind of trials, time to event endpoints are often chosen as primary endpoints.

Conclusions about phase II trials are made through a hypothesis testing procedure that, for instance, has the form

$$H_0 : p \leq p_0 \text{ vs. } H_1 : p \geq p_1, \tag{1.7}$$

where the type-I error is usually increased from the traditional 5% up to 20% in some particular cases. However,the type-II error needs to be bounded between 10% and 20%. The reason is that in phase II trials the main goal is identify active treatments and hence controlling the type-II error is slightly more important than controlling the type-I error.

Given the randomized nature of phase IIB trials, it is not rare to find trials that use adaptive randomization. The reason is that with this approach, we can allocate patients to more promising treatment while still having a randomized setting in the clinical trial.

One the most important articles published so far is the two-stage design proposed by [15].

1.3 Confirmatory clinical trials

In this section we briefly introduce adaptive phase III (or confirmatory) clinical trials. They are usually randomized, controlled, multi-center, multi-arm and multi-stage trials. They compare the dose recommended by the phase II against the current *gold standard* treatment and enroll a much larger number of patients than the previous stages of the drug development.

What differentiates a confirmatory trial from an adaptive confirmatory trial is the possibility of looking at the data several times while the trial is ongoing for interim decision making. The rationale for this is that a clinical trial should not continue if there is a clear tendency favoring one of the treatments. Also, patients should not be treated with a drug that does not show a potential benefit for them.

One of the biggest challenges in adaptive confirmatory trials is the type-I error rate control as it increases if we repeatedly look at the data. Methods to prevent type-I error rate inflation are hence necessary in order to obtain valid conclusions from these kind of clinical trials. These method were introduced by [16–18], amont others, are widely used in the implementations of group sequential testing for clinical trials.

As described by [19], there are two kind of adaptive confirmatory trials: group sequential designs, which are characterized by a pre-specified adaptivity, and confirmatory adaptive designs, which are characterized by unscheduled adaptivity.

In group sequential designs, the number of interim analysis, the sample size of each arm or the decision boundaries for early stopping need to be fixed before the trial has started to avoid a type-I error rate inflation. In contrast, confirmatory adaptive trials allow to perform any adaptivity, such us for example sample size re-assessment, without a pre-specification of the adaption rules. To prevent type-I error rate inflation in adaptive confirmatory trials, the design must satisfy the *conditional invariance principle* (see [20]), where different test statistics are calculated from the samples at the different stages of the trial and are combined in a pre-specified way for the final test decisions. This principle allows to react quickly to unexpected results.

Chapter 2

Cancer Phase I trial design using drug combinations when a fraction of dose limiting toxicities is attributable to one or more agents

2.1 Introduction

Cancer phase I clinical trials constitute the first step in investigating a potentially promising combination of cytotoxic and biological agents. Due to safety and ethical concerns, patients are sequentially enrolled in the trial, and the dose combinations assigned to subsequent patients depend on dose combinations already given to previous patients and their dose limiting toxicity (DLT) status at the end of the first cycle of therapy. The main objective of these trials is to estimate a maximum tolerated dose (MTD) that will be used in future efficacy evaluation in phase II/III trials. The MTD is usually defined as any dose combination (x,y) that will produce DLT in a prespecified proportion θ of patients,

$$\text{Prob}(\text{DLT}|\ \text{dose} = (x,y)) = \theta. \tag{2.1}$$

The definition of DLT depends on the type of cancer and drugs under study, but it is usually defined as a grade 3 or 4 non-hematologic toxicity (see the National Cancer Institute CTCAE v4.03 for the definition of the different grades of toxicity). The pre-specified

proportion of DLTs θ, sometimes referred as target probability of DLT, also depends on the nature of the toxicity, but it usually take values between 0.2 and 0.4.

In the drug combination dose finding literature, designs that recommend a unique MTD (see e.g. [6, 7, 5, 8–10, 21, 22]) or multiple MTDs (see e.g. [23–27, 12, 28, 13]) have been studied extensively. Most of these methods use a parametric model for the dose-toxicity relationship

$$\text{Prob}(\text{DLT}|(x,y)) = F((x,y),\xi), \tag{2.2}$$

where (x,y) represents the drug combination of two agents, $F(.)$ is a known link function, e.g. a power model or a logistic model, and $\xi \in R^d$ is a vector of d unknown parameters. Non-parametric designs have been proposed in the past, both in single agent and drug combination settings [27, 29, 30]. These designs unique assumption is monotonicity, which is imposed either through the prior distribution (see [29, 30]), or by choosing only monotonic contours when escalating (see [27]).

Let S be the set of all dose combinations available in the trial, and $C(\xi)$ be the set of dose combinations (x,y) such the probability of DLT equals a target risk of toxicity θ. Hence,

$$C(\xi) = \{(x,y) \in S : F((x,y),\xi) = \theta\}. \tag{2.3}$$

Equation (2.3) is the traditional definition of MTD set. When S is discrete, following [13], we can define the MTD as the set of dose combinations (x,y) that satisfy

$$|F((x,y),\xi) - \theta| \leq \delta, \tag{2.4}$$

since $C(\xi)$ may be empty, i.e., when the MTD is not in S. The threshold parameter $\delta, 0 < \delta < 1$, is pre-specified after close collaboration with the clinician.

This work is motivated by a cancer phase I trial a clinician at Cedars-Sinai Medical Center is planning. The trial involves the combination of Taxotere, a known cytotoxic agent, and Metformin, a diabetes drug, in advanced or metastatic breast cancer patients. According to the clinician, some DLTs can be attributable to either agent or both. For example, a grade 3 or 4 neutropenia can only be attributable to Taxotere and not Metformin. Furthermore, for ethical reasons, if a patient has a DLT attributable to Taxotere when treated with dose level x_T of taxotere, then x_T cannot be increased for the next patient in the trial (see the dose escalation restriction in Section 2.2). Very few methods have been developed to incorporate toxicity attribution in the dose escalation process. [6] proposed a design that models the

joint probability of toxicity with a copula model known as the Gumbel model [31]. This model allows the investigator to compute the probability of DLT when the DLT is exclusively attributed to one drug, the other one, or both. However, they require all toxicities to be attributable, which is rare in practice. [21] proposed a semi-attributable toxicity design based on a trial with non-concurrent drug administration. In their design, one drug is administered at the beginning of the treatment cycle and the other drug is administered at a much later time point if and only if the patient did not experience DLT. If a DLT occurs before the second drug is administered, then the DLT is attributed to the first drug. However, if the DLT occurs after the second drug has been administered, then the DLT could be caused by any of the drugs and therefore is not attributable. [32] propose a method that reduces the effect of the bias caused by toxicity attribution errors by using personalized scores instead of the traditional binary DLT outcome. [33] considered the toxicity attribution problem for ruled-based designs with non-overlapping toxicities.

In this article, we propose a Bayesian adaptive design for drug combinations that allows the investigator to attribute a DLT to one or both agents in an unknown fraction of patients, even when the drugs are given concurrently.

We define toxicity attribution as a DLT caused by one drug and not the other when the type of DLT is non-overlapping, e.g., a grade 4 neutropenia is caused by taxotere but can never occur with metformin, or when the clinician judges that a type of DLT is caused by one drug and not the other, e.g., a grade 4 diarrhea is caused by taxotere but not metformin due to the low dosTe level of taxotere that was given in combination even though both drugs have this side effect in common.

The relationship between the dose combinations and the risk of toxicity is modeled using the same copula model used by [6]. The design proceeds using a variation of the algorithm proposed in [13] where cohorts of two patients are allocated to dose combinations where, at each stage of the trial, we search for the dose of one agent given the current dose of the other agent. Our approach differs from the methodologies of [6] and [13] in three aspects; (i) a non-negative fraction of DLTs are attributable to either one or both agents, (ii) the dose combination allocated to patients uses the CRM scheme as opposed to escalation with overdose control (EWOC) approach proposed by [11], and (iii) if a current patient experiences DLT attributed drug D_1 at dose level x_{D_1}, then the dose level of agent D_1 cannot be more than x_{D_1} for the next cohort of two patients. At the end of the trial, an estimate of the MTD curve is proposed as a function of Bayes estimates of the model parameters. Last, we show that our method can be easily adapted from a setting with continuous dose

combinations to discrete dose combinations by rounding up the estimated MTD curve to the nearest discrete dose combinations.

The rest of the manuscript is organized as follows. In Section 2.2, we describe the model for the dose-toxicity relationship and the adaptive design to conduct the trial for continuous dose combinations. In Section 2.3, we study the performance of the method in terms of safety and efficiency of the estimate of the MTD set. In Section 2.4, we adapt our proposal to the setting of discrete dose combinations. In section 2.5, we conduct a model misspecification evaluation. Discussion and practical considerations of the method are discussed in Section 2.6.

2.2 Method

2.2.1 Dose-Toxicity Model

Let $X_{\min}, X_{\max}, Y_{\min}, Y_{\max}$, be the minimum and maximum doses available in a trial that combines drugs with continuous dose combination levels. The doses are standardized to be in a desired interval, e.g., [0.05, 0.3], so that $X_{\min} = Y_{\min} = 0.05$ and $X_{\max} = Y_{\max} = 0.3$. Let $F_\alpha(\cdot)$ and $F_\beta(\cdot)$ be parametric models for the probability of DLT of drugs D_1 and D_2, respectively. We specify the joint dose-toxicity relationship using the Gumbel copula model (see [31]) as

$$\begin{aligned}\pi^{(\delta_1,\delta_2)} &= \text{Prob}(\delta_1,\delta_2|x,y) = F_\alpha^{\delta_1}(x)\left[1-F_\alpha(x)\right]^{1-\delta_1} \times \\ &F_\beta^{\delta_2}(y)\left[1-F_\beta(y)\right]^{1-\delta_2} + (-1)^{(\delta_1+\delta_2)}F_\alpha(x)\left[1-F_\alpha(x)\right]F_\beta(y)\left[1-F_\beta(y)\right]\frac{e^{-\gamma}-1}{e^{-\gamma}+1},\end{aligned} \tag{2.5}$$

where x is the standardized dose level of drug D_1, y is the standardized dose level of agent D_2, δ_1 is the binary indicator of DLT attributed to drug D_1, δ_2 is the binary indicator of DLT attributed to drug D_2 and γ is the interaction coefficient. We assume that the joint probability of DLT, when one of the drugs is held constant, is monotonically increasing; that is $\text{Prob}(\text{DLT}|x',y) \geq \text{Prob}(\text{DLT}|x,y)$ or $\text{Prob}(\text{DLT}|x,y') \geq \text{Prob}(\text{DLT}|x,y)$, where $x' > x$ and $y' > y$. A sufficient condition for this property to hold is to assume that $F_\alpha(\cdot)$ and $F_\beta(\cdot)$ are increasing functions with $\alpha > 0$ and $\beta > 0$. In this article we use $F_\alpha(x) = x^\alpha$ and $F_\beta(y) = y^\beta$. Using (2.5), if the DLT is attributed exclusively to drug D_1, then

$$\pi^{(\delta_1=1,\delta_2=0)} = \text{Prob}(\delta_1 = 1, \delta_2 = 0|x,y) = x^{\alpha}(1-y^{\beta}) - x^{\alpha}(1-x^{\alpha})y^{\beta}\left(1-y^{\beta}\right)\frac{e^{-\gamma}-1}{e^{-\gamma}+1}. \tag{2.6}$$

If the DLT is attributed exclusively to drug D_2, then

$$\pi^{(\delta_1=0,\delta_2=1)} = \text{Prob}(\delta_1 = 0, \delta_2 = 1|x,y) = y^{\beta}(1-x^{\alpha}) - x^{\alpha}(1-x^{\alpha})y^{\beta}\left(1-y^{\beta}\right)\frac{e^{-\gamma}-1}{e^{-\gamma}+1}. \tag{2.7}$$

If the DLT is attributed to both drugs D_1 and D_2, then

$$\pi^{(\delta_1=1,\delta_2=1)} = \text{Prob}(\delta_1 = 1, \delta_2 = 1|x,y) = x^{\alpha}y^{\beta} + x^{\alpha}(1-x^{\alpha})y^{\beta}\left(1-y^{\beta}\right)\frac{e^{-\gamma}-1}{e^{-\gamma}+1}. \tag{2.8}$$

Equation (2.6) represents the probability that D_1 causes a DLT and drug D_2 does not cause a DLT. This can happen, for example, when a type of DLT of taxotere (D_1), such as grade 4 neutropenia, is observed. However, this type of DLT can never be observed with metformin (D_2). This can also happen when the clinician attributes a grade 4 diarrhea to taxotere (D_1) but not to metformin (D_2) in the case of a low dose level of this later even though both drugs have this common type of side effect. The fact that dose level y is present in equation (2.6) is a result of the joint modeling of the two marginals and accounts for the probability that drug D_2 does not cause a DLT. This later case is, of course, based on the clinician's judgment. Equations (2.7) and (2.8) can be interpreted similarly.

Following [6], it is easy to see that the total probability of having a DLT is calculated as the sum of (2.6), (2.7) and (2.8). Hence,

$$\begin{aligned}\pi = \text{Prob}(\text{DLT}|x,y) = \pi^{(\delta_1=1,\delta_2=0)} + \pi^{(\delta_1=0,\delta_2=1)} + \pi^{(\delta_1=1,\delta_2=1)} = \\ x^{\alpha} + y^{\beta} - x^{\alpha}y^{\beta} - x^{\alpha}(1-x^{\alpha})y^{\beta}\left(1-y^{\beta}\right)\frac{e^{-\gamma}-1}{e^{-\gamma}+1}.\end{aligned} \tag{2.9}$$

We define the MTD as any dose combination (x_*, y_*) such that Prob(DLT$|x_*, y_*) = \theta$. We set (2.9) equal to θ and re-write it as a 2nd degree polynomial in y^β, and solve for the solutions. This allows us to define the MTD set $C(\alpha, \beta, \gamma)$ as

$$C(\alpha,\beta,\gamma) = \left\{ (x_*, y_*) : y_* = \left[\frac{-(1 - x_*^\alpha - \kappa) \pm \sqrt{(1 - x_*^\alpha - \kappa)^2 - 4\kappa(x_*^\alpha - \theta)}}{2\kappa} \right]^{\frac{1}{\beta}} \right\}, \tag{2.10}$$

where

$$\kappa = x_*^\alpha (1 - x_*^\alpha) \frac{e^{-\gamma} - 1}{e^{-\gamma} + 1}.$$

Let T be the indicator of DLT, $T = 1$ if a patient treated at dose combination (x, y) experiences DLT within one cycle of therapy that is due to either drug or both, and $T = 0$ otherwise. Among patients treated with dose combination (x, y) who exhibit DLT, suppose that an unknown fraction η of these patients have a DLT with known attribution, i.e. the clinician knows if the DLT is caused by drug D_1 only, or drug D_2 only, or both drugs D_1 and D_2. Let A be the indicator of DLT attribution when $T = 1$. It follows that for each patient treated with dose combination (x, y), there are five possible toxicity outcomes: $\{T = 0\}, \{T = 1, A = 0\}, \{T = 1, A = 1, \delta_1 = 1, \delta_2 = 0\}, \{T = 1, A = 1, \delta_1 = 0, \delta_2 = 1\}$ and $\{T = 1, A = 1, \delta_1 = 1, \delta_2 = 1\}$. This is illustrated in the chance tree diagram in Figure 2.1. Using equations (2.6),(2.7),(2.8),(2.9) and Figure 2.1, the contributions to the likelihood from each of the five observable outcomes are listed in Table 2.1. The likelihood function is defined as

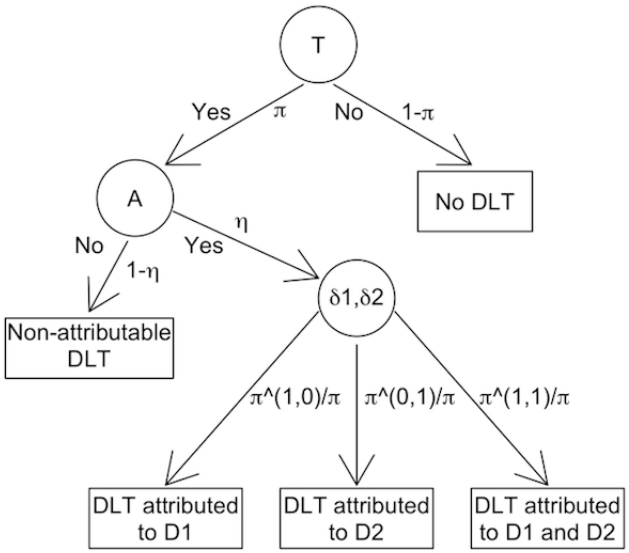
$$L(\alpha,\beta,\gamma,\eta \mid \text{data}) = \prod_{i=1}^{n} [(\eta \pi_i^{(\delta_{1_i}, \delta_{2_i})})^{A_i} (\pi_i \, (1 - \eta))^{1 - A_i}]^{T_i} (1 - \pi_i)^{1 - T_i}, \tag{2.11}$$

and the joint posterior probability distribution of the model parameters as

$$\text{Prob}(\alpha,\beta,\gamma,\eta \mid \text{data}) \propto \text{Prob}(\alpha,\beta,\gamma) \times L(\alpha,\beta,\gamma \mid \text{data}). \tag{2.12}$$

With equation (2.12) we can easily sample and obtain MCMC estimates of α, β, γ and η.

Fig. 2.1 A chance tree illustrating the 5 possible outcomes we can find in a trial.



2.2.2 Trial Design

Dose escalation / de-escalation proceeds using the algorithm described in [13] but univariate continual reassessment method (CRM) is carried out to estimate the next dose instead of EWOC. In a cohort with two patients, the first one would receive a new dose of agent D_1 given the dose y of agent D_2 that was previously assigned. The new dose of agent D_1 is defined as $x_{\text{new}} = \text{argmin}_u|\widehat{\text{Prob}}(\text{DLT}|u,y) - \theta|$,where y is fixed and $\hat{\text{Prob}}(\text{DLT}|u,y)$ is computed using equation (2.9) with α, β, γ replaced by their posterior medians. The other patient would receive a new dose of agent D_2 given the dose of agent D_1 that was previously assigned. Specifically, the design proceeds as follows:

1. Patients in the first cohort receive the same dose combination $(x_1, y_1) = (x_2, y_2) = (X_{\min}, Y_{\min})$.

2. In the i-th cohort of two patients,

   - If i is even,
     - Patient $(2i-1)$ receives doses (x_{2i-1}, y_{2i-1}), where $x_{2i-1} = \underset{u}{\text{argmin}} \left|\widehat{\text{Prob}}(\text{DLT}|u, y_{2i-3}) - \theta\right|$, and $y_{2i-1} = y_{2i-3}$. If a DLT was observed

Table 2.1 Contributions to the likelihood function based on the observed outcomes: toxicity, attribution, attribution to drug 1 (δ_1) and attribution to drug 2 (δ_2) for each patient.

Toxicity	Attribution	δ_1	δ_2	Likelihood
0	-	-	-	$1-\pi = 1-[x^\alpha + y^\beta - x^\alpha \times y^\beta - x^\alpha(1-x^\alpha)y^\beta\left(1-y^\beta\right)\frac{e^{-\gamma}-1}{e^{-\gamma}+1}]$
1	0	-	-	$\pi\times(1-\eta) = [x^\alpha + y^\beta - x^\alpha \times y^\beta - x^\alpha(1-x^\alpha)y^\beta\left(1-y^\beta\right)\frac{e^{-\gamma}-1}{e^{-\gamma}+1}]\times(1-\eta)$
1	1	1	0	$\pi\times\eta\times\frac{\pi^{(1,0)}}{\pi} = \eta\times[x^\alpha(1-y^\beta) - x^\alpha(1-x^\alpha)y^\beta\left(1-y^\beta\right)\frac{e^{-\gamma}-1}{e^{-\gamma}+1}]$
1	1	0	1	$\pi\times\eta\times\frac{\pi^{(0,1)}}{\pi} = \eta\times[y^\beta(1-x^\alpha) - x^\alpha(1-x^\alpha)y^\beta\left(1-y^\beta\right)\frac{e^{-\gamma}-1}{e^{-\gamma}+1}]$
1	1	1	1	$\pi\times\eta\times\frac{\pi^{(1,1)}}{\pi} = \eta\times[x^\alpha\times y^\beta + x^\alpha(1-x^\alpha)y^\beta\left(1-y^\beta\right)\frac{e^{-\gamma}-1}{e^{-\gamma}+1}]$

in the previous cohort of two patients and was attributable to drug D_1, then x_{2i-1} is further restricted to be no more than x_{2i-3}.

- Patient $2i$ receives doses (x_{2i}, y_{2i}), where $y_{2i} = \underset{v}{\text{argmin}}\left|\widehat{\text{Prob}}(\text{DLT}|x_{2i-2}, v) - \theta\right|$, and $x_{2i} = x_{2i-2}$. If a DLT was observed in the previous cohort of two patients and was attributable to drug D_2, then y_{2i} is further restricted to be no more than y_{2i-2}.

- If i is odd,
  - Patient $(2i-1)$ receives doses (x_{2i-1}, y_{2i-1}), where $y_{2i-1} = \underset{v}{\text{argmin}}\left|\widehat{\text{Prob}}(\text{DLT}|x_{2i-3}, v) - \theta\right|$, and $x_{2i-1} = x_{2i-3}$. If a DLT was observed in the previous cohort of two patients and was attributable to drug D_2, then y_{2i-1} is further restricted to be no more than y_{2i-3}.
  - Patient $2i$ receives doses (x_{2i}, y_{2i}), where $x_{2i} = \underset{u}{\text{argmin}}\left|\widehat{\text{Prob}}(\text{DLT}|u, y_{2i-2}) - \theta\right|$, and $y_{2i} = y_{2i-2}$. If a DLT was observed in the previous cohort of two patients and was attributable to drug D_1, then x_{2i} is further restricted to be no more than x_{2i-2}.

3. Repeat step 2 until the maximum sample size is reached subject to the following stopping rule.

4. We would stop the trial if, $\text{Prob}(\text{Prob}(\text{DLT}|x = X_{\min}, y = Y_{\min}) \geq \theta + \xi_1|\text{data}) > \xi_2$, i.e. if the posterior risk of toxicity at the lowest combination significantly is high. ξ_1 and ξ_2 are design parameters tuned to obtain the best operating characteristics.

In step 2 of the algorithm, any dose escalation is further restricted to be no more than a pre-specified fraction of the dose range of the corresponding agent. At the end of the trial, we

obtain the MTD curve estimate $\hat{C} = C(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$, where $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ are the posterior medians of the parameters α, β and γ, given the data.

2.3 Simulation Studies

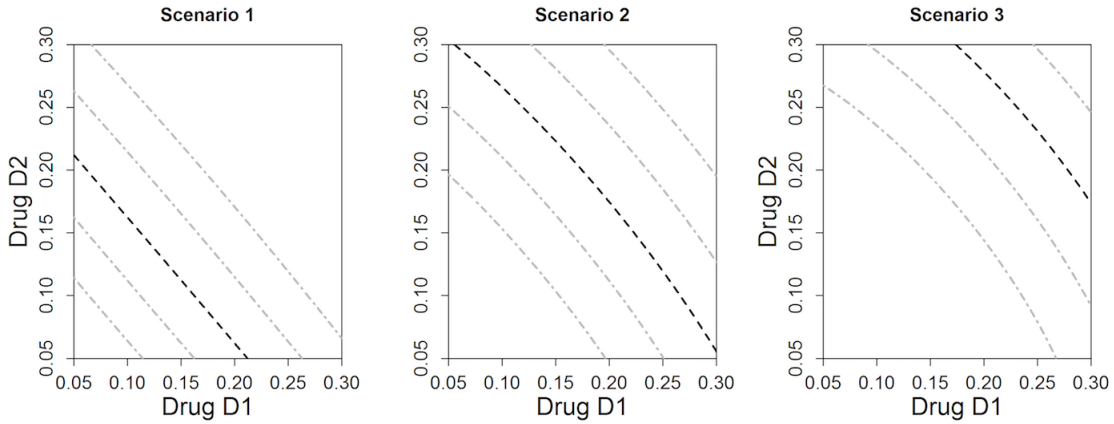
2.3.1 Simulation set up and Scenarios

In all simulated trials, the link functions $F_\alpha(x) = x^\alpha$ and $F_\beta(y) = y^\beta$ are used. To evaluate the performance of our proposal, the DLT outcomes are generated from the true model showed in (2.9). We used this model in 3 different scenarios to study the behavior of our design when the prior distribution of the model parameters is both well and poorly calibrated. Let α_{true}, β_{true} and γ_{true} represent the true parameter values we use in (2.9) to generate DLT outcomes. In each scenario we select different values for α_{true}, β_{true}, but the prior distribution for α and β, $P(\alpha)$ and $P(\beta)$, as well as γ_{true}, do not vary. In scenario 1, we choose values for α_{true} and β_{true} such that $\alpha_{\text{true}} < E[P(\alpha)]$ and $\beta_{\text{true}} < E[P(\beta)]$. In scenario 2, we choose values for α_{true} and β_{true} such that $\alpha_{\text{true}} = E[P(\alpha)]$ and $\beta_{\text{true}} = E[P(\beta)]$. Last, in scenario 3, we choose values for α_{true} and β_{true} such that $\alpha_{\text{true}} > E[P(\alpha)]$ and $\beta_{\text{true}} > E[P(\beta)]$. Figure 2.2 shows the MTD curves with the true parameter values described here and their contours at $\theta \pm 0.05$ and $\theta \pm 0.1$. We evaluate the effect of toxicity attribution in these 3 scenarios using 4 different values for η: 0, 0.1, 0.25 and 0.4. These values are reasonable because higher values of η in practice are very rare. Data is randomly generated using the following procedure:

- For a given dose combination (x, y), a binary indicator of DLT T is generated from a Bernoulli distribution with probability of success computed using equation (2.9).
- If $\{T = 1\}$, we generate the attribution outcome A using a Bernoulli distribution with probability of success η.
- If $\{T = 1, A = 1\}$, we attribute the DLT to drug D_1, D_2, or to both drugs with equal probabilities.

We assume that the model parameters α, β, γ and η are independent *a priori*. We assign vague prior distributions to α, β and γ following [6], where $\alpha \sim \text{Uniform}(0.2, 2)$, $\beta \sim \text{Uniform}(0.2, 2)$ and $\gamma \sim \text{Gamma}(0.1, 0.1)$. These prior distributions correspond to the ones used by [6] for the main analysis. The prior distribution for the fraction of attributable toxicities η is set to be $\text{Uniform}(0, 1)$. With these prior distributions, the true parameter

Fig. 2.2 Contour plots for the working model in scenarios 1, 2 and 3. The black dashed curve represents the true MTD curve and the gray dashed lines represent the contours at $\theta \pm 0.05$ and $\theta \pm 0.10$.



values for each scenario are as follows. In scenario 1, $\alpha = \beta = 0.9$ and $\gamma = 1$. In scenario 2, $\alpha = \beta = 1.1$ and $\gamma = 1$. Last, in scenario 3, $\alpha = \beta = 1.3$ and $\gamma = 1$. For each scenario, $m = 1000$ trials will be simulated. The target risk of toxicity is fixed at $\theta = 0.3$, the sample size is $n = 40$, and the values for ξ_1 and ξ_2 will be 0.05 and 0.8 respectively. All simulation are done using the software R version 3.3.1.

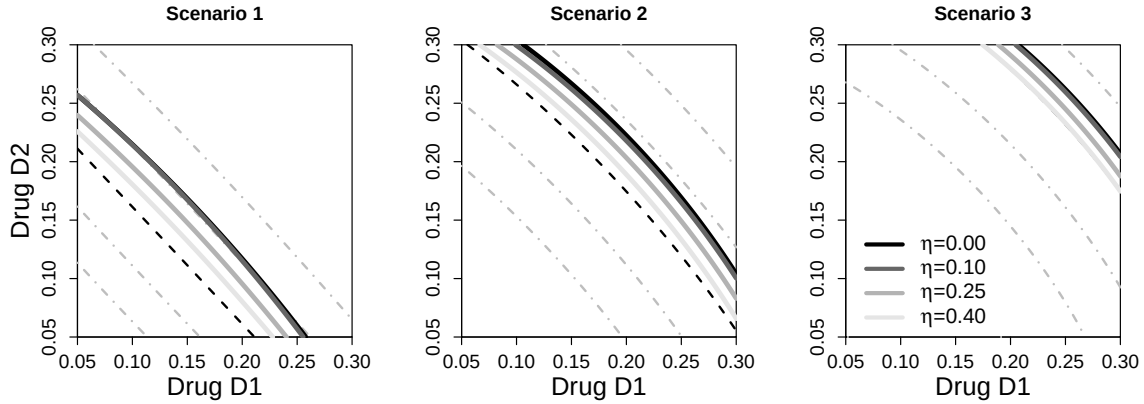
2.3.2 Design Operating Characteristics

We evaluate the performance of the design by assessing its safety and its efficiency in estimating the MTD curve.

For trial safety, we employ the average percent of DLTs, the percent of simulated trials with DLT rate greater than $\theta \pm 0.05$ and $\theta \pm 0.10$.

For efficiency, we employ the pointwise average relative minimum distance from the true MTD curve to the estimated MTD curve. This measure of efficiency is well described in [12, 13] and can interpreted as a pointwise average bias in estimating the true MTD curve. We also consider the pointwise percent of trials for which the minimum distance of the point (x,y) on the true MTD curve to the estimated MTD curve is no more than $(100 \times p)\%$ of the true MTD curve. This measurement will give us an estimate of the percent of trials with MTD recommendation within $(100 \times \text{p})\%$ of the true MTD. This measure of efficiency can be interpreted as the pointwise percent of correct MTD recommendation. In this paper we select $p = 0.1, 0.2$. For a detailed explanation of these measures of efficiency, see [12, 13].

Fig. 2.3 Estimated MTD curves for $m = 1000$ simulated trials. The black dashed curve represents the true MTD curve, the gray dashed lines represent the contours at $\theta \pm 0.05$ and $\theta \pm 0.10$, and the solid curves represent the estimated MTD curves at each value of η.



2.3.3 Results

In general, increasing the value of η until 0.4 generates estimated MTD curves closer to the true MTD curve. Figure 2.3 shows the estimated MTD curves for each scenario as a function of η. In terms of safety, overall we observe that increasing the fraction of toxicity attributions η reduces the average percent of toxicities and percent of trials with toxicity rates greater than $\theta + 0.05$ and $\theta + 0.10$. Table 2.2, shows the average percent of toxicities as well as the percent of trials with toxicity rates greater than $\theta + 0.05$ and $\theta + 0.1$ for scenarios 1-3.

Figure 2.4 shows the pointwise average bias of the 3 proposed scenarios for each value of η. Overall, increasing the value of η until 0.4 reduces the pointwise average bias. In any case, the pointwise average bias is around 10% of the dose range of either drug and practically negligible for $\eta = 0.25, 0.4$. For instance, under scenario 3, the maximum absolute value of the pointwise average bias when $\eta = 0.40$ is about 0.01, which corresponds to 0.3% of the dose range, which is practically negligible.

Figure 2.5 shows the pointwise percent of MTD recommendation of the 3 proposed scenarios for each value of η. In general, increasing the value of η increases the pointwise percent of MTD recommendation, reaching up to 80% of correct recommendation when $p = 0.2$, and up to 70% of correct recommendation when $p = 0.1$. Based on these simulation results, we conclude that in continuous dose setting the approach of partial toxicity attribution generates safe trial designs and efficient estimation of the MTD.

Fig. 2.4 Pointwise average bias in estimating the true MTD in $m = 1000$ simulated trials.

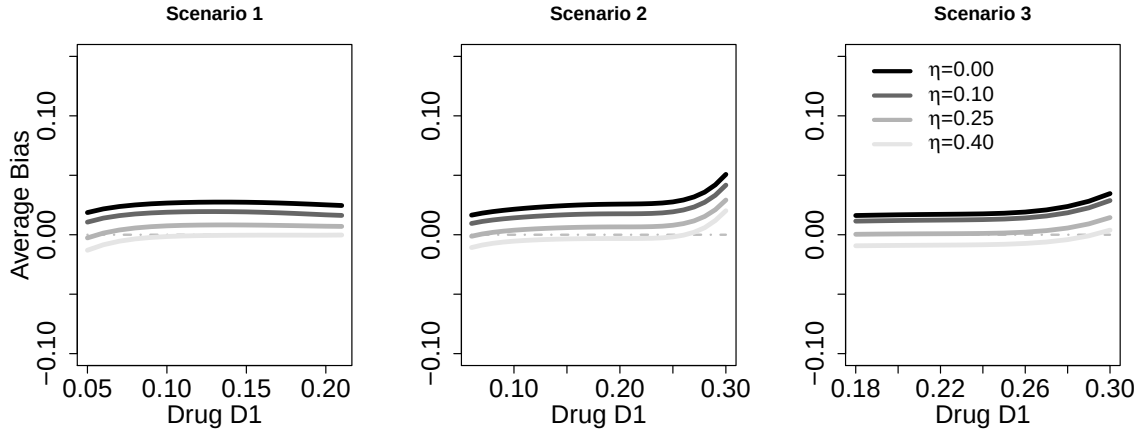
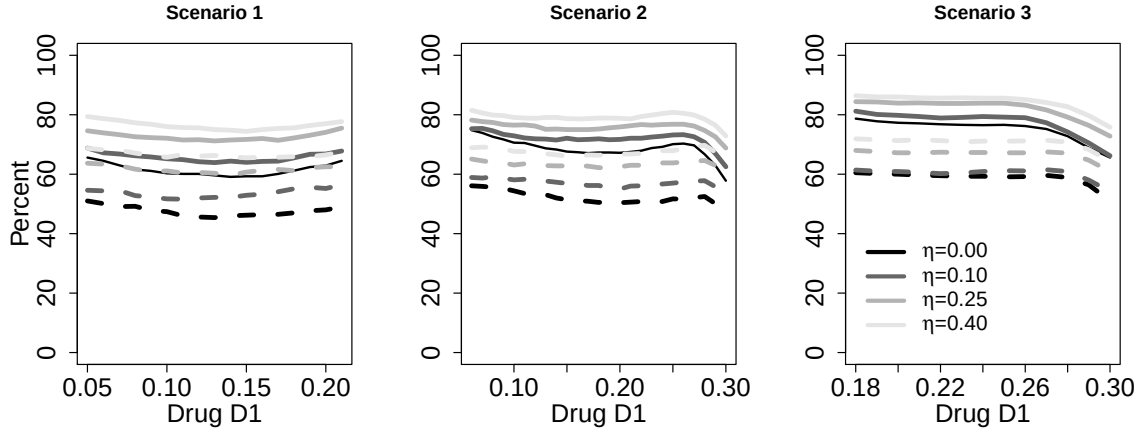


Fig. 2.5 Pointwise percent of MTD recommendation for $m = 1000$ simulated trials. Solid lines represent the pointwise percent of MTD recommendation when $p = 0.2$ and dashed lines represent the pointwise percent of MTD recommendation when $p = 0.1$.



2.4 Discrete Dose Combinations

2.4.1 Approach

Dose escalation follows the same procedure described in section 2.2.2. The only difference is that, in step 2, the continuous doses recommended are rounded to the nearest discrete dose level. For a detailed explanation of this procedure see [13].

Table 2.2 Operating characteristics summarizing trial safety in $m = 1000$ simulated trials.

		Average % of toxicities	% of trials with toxicity rate $> \theta + 0.05$	% of trials with toxicity rate $> \theta + 0.10$
Scenario 1	$\eta = 0.00$	33.62	25.90	4.10
	$\eta = 0.10$	32.67	22.60	4.80
	$\eta = 0.25$	31.55	17.60	2.70
	$\eta = 0.40$	30.70	13.30	2.00
Scenario 2	$\eta = 0.00$	30.64	9.40	0.90
	$\eta = 0.10$	29.69	7.30	0.40
	$\eta = 0.25$	28.76	5.00	0.20
	$\eta = 0.40$	28.04	4.10	0.30
Scenario 3	$\eta = 0.00$	27.47	2.00	0.00
	$\eta = 0.10$	26.80	1.80	0.00
	$\eta = 0.25$	25.99	1.30	0.00
	$\eta = 0.40$	25.37	0.70	0.00

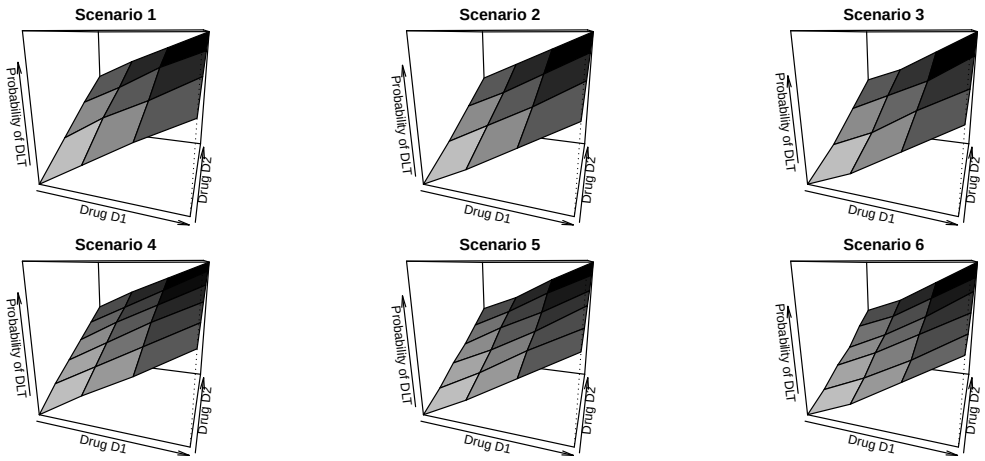
2.4.2 Illustration

We study the performance of our proposal in a discrete dose level setting where the probability of toxicity of each dose level is generated from the working model. We employ 6 scenarios with 4 dose levels respectively in each drug for scenarios 1 - 3, and 4 and 6 dose levels respectively in each drug for scenarios 4 - 6. The target probability of toxicity is always $\theta = 0.3$ and, for each scenario, we simulate $m = 1000$ trials using the same vague priors for α, β and γ specified in section 2.3.1. The maximum sample size in all scenarios is again $n = 40$. The performance of the method is evaluated using the percent of MTD selection statistic proposed by [13].

In Table 2.3 we present the 6 mentioned scenarios we use to illustrate the implementation of our design with discrete dose levels. Moreover, in Figure 2.6 we show the dose-toxicity surface of these 6 scenarios, where we observe that all of them have a flat (near-constant) surface.

In Table 2.4 we show the percent of times that at least 25%, 50%, 75% or 100% of recommended MTDs belong to the true MTD set. Using vague prior distributions, the scenario where toxicity attribution has the strongest effect is scenario 2. In scenarios 1,4 and 5, we observe a slight effect but it does not make a big difference.

Fig. 2.6 Probability of DLT surfaces of the 6 scenarios from Table 2.3.



2.5 Model Misspecification

In the previous sections, all the simulated scenarios are generated with the model showed in (2.9). However, in practice we do not know the underlying model that generates the data and therefore we need to assess the performance of our design under model misspecification. We employ the same toxicity scenarios used by [6], which are shown in Table 2.5. Moreover, In Figure 2.7 we show the dose-toxicity surface of these scenarios. Scenario 1 presents a very constant surface gradient. The rest of the scenarios present surface gradients that vary as we increase the dose combination levels. However, scenarios 3, 4 and 6 vary more abruptly than scenarios 2 and 5. Scenario 6 is a particular case because the lowest dose combination level has a probability of DLT that is already higher than the target risk of toxicity $\theta + 0.1$. Therefore, for this scenario, instead of presenting the percent of correct recommendation we present the percent of times the trial is stopped due to safety using the stopping rule in Section 2.2 with $\xi_1 = 0.05$ and $\xi_2 = 0.8$. For each scenario, we simulate $m = 1000$ trials with a target risk of toxicity of $\theta = 0.30$, a sample size of $n = 40$ and we use the same prior distributions for α, β and γ as in section 2.3.1.

In terms of safety, in general we observe that toxicity attributions reduce the average percent of toxicities and percent of trials with toxicity rates greater than $\theta + 0.05$ and $\theta + 0.10$. Table 2.6 shows the average percent of toxicities as well as the percent of trials with toxicity rates greater than $\theta + 0.05$ and $\theta + 0.1$.

In Table 2.7, we show the percent of times that at least 25%, 50%, 75% or 100% of recommended MTDs belong to the true MTD set. In scenario 1 we observe a positive effect of the toxicity attributions, improving the percent of times at least 75% and 100% of

Table 2.3 Dose limiting toxicity scenarios with $\theta = 0.3$ generated with the working model. In bold the dose combination levels that would compose the true MTD set.

Dose level	1	2	3	4	1	2	3	4	5	6
	Scenario 1				Scenario 4					
4	**0.39**	0.46	0.52	0.58	**0.39**	0.43	0.47	0.51	0.55	0.58
3	**0.31**	**0.38**	0.46	0.52	**0.30**	**0.35**	0.40	0.44	0.48	0.52
2	**0.22**	**0.31**	**0.38**	0.46	**0.22**	**0.27**	**0.32**	**0.37**	0.41	0.46
1	0.13	**0.22**	**0.31**	**0.39**	0.13	0.19	**0.24**	**0.29**	**0.34**	**0.39**
	Scenario 2				Scenario 5					
4	**0.30**	**0.36**	0.42	0.48	**0.30**	**0.33**	**0.37**	0.40	0.44	0.48
3	**0.22**	**0.28**	**0.35**	0.42	**0.22**	**0.26**	**0.29**	**0.33**	**0.38**	0.42
2	0.14	**0.21**	**0.28**	**0.36**	0.14	0.18	**0.22**	**0.27**	**0.31**	**0.35**
1	0.07	0.14	**0.22**	**0.30**	0.07	0.11	0.16	0.20	**0.25**	**0.30**
	Scenario 3				Scenario 6					
4	**0.23**	**0.27**	**0.33**	**0.39**	**0.23**	**0.25**	**0.28**	**0.32**	**0.35**	**0.39**
3	0.16	**0.21**	**0.26**	**0.33**	0.16	0.18	**0.22**	**0.25**	**0.29**	**0.33**
2	0.09	0.14	**0.21**	**0.27**	0.09	0.12	0.16	0.19	**0.23**	**0.27**
1	0.04	0.09	0.16	**0.23**	0.04	0.07	0.11	0.14	0.19	**0.23**

recommended MTDs belong to the true MTD set in to 5% when $\eta = 0.25$. In scenario 2 we observe a positive effect of the toxicity attributions improving the percent of times at least 25% and 50% of recommended MTDs belong to the true MTD up to 5% and 4% respectively. In scenarios 3, 4 and 5 we do not observe any positive effect when attributing toxicities. However, these scenarios are particularly difficult for our design given the rounding up procedure we follow with discrete dose combinations. In scenario 6 we observe a positive effect of the toxicity attributions, improving the percent of times the trial is stopped due to safety by almost 4% when $\eta = 0.40$. Based on these simulation results under model misspecification, we conclude that the partial toxicity attribution method has good operating characteristics in recommending dose combinations of which, at least 50% are the true MTDs; these percent of correct recommendations vary between 65% to 98% depending on the scenario. Moreover, there is a high probability of stopping the trial if there is evidence that the minimum dose combination in the trial has high probability of DLT.

However some of the scenarios showed in Table 2.5 have a true set of MTDs that include a large number of dose combinations. For this reason, we implemented our design in 6 extra scenarios taken from [6, 7]. These scenarios are presented in Table 2.8 at the supplementary material, where the set of true MTDs contains a much more restricted number of dose

Table 2.4 Percent of times that at least 25%, 50%, 75% or 100% of recommended MTDs belong to the true MTD set in $m = 1000$ simulated trials.

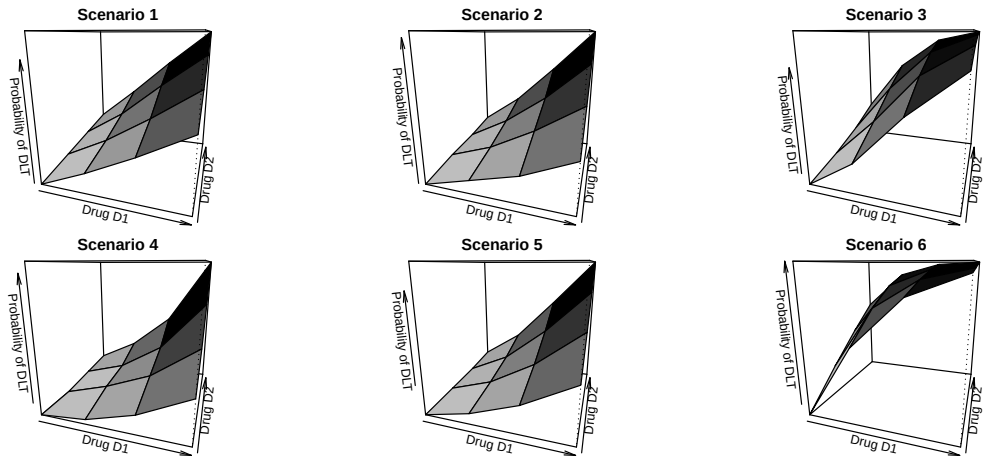
		% of correct MTD recommendation for $\theta \pm 0.10$							
		$\geq 25\%$	$\geq 50\%$	$\geq 75\%$	100%		$\geq 25\%$	$\geq 50\%$	$\geq 75\%$	100%
$\eta = 0.00$		91.40	87.30	83.70	83.70		90.00	85.50	**73.70**	67.70
$\eta = 0.10$		**92.50**	**87.80**	**83.90**	**83.90**		**91.30**	**86.00**	71.10	65.80
$\eta = 0.25$	Scenario 1	90.90	87.70	83.80	83.80	Scenario 4	89.40	85.90	71.50	**68.20**
$\eta = 0.40$		90.90	87.70	83.80	83.80		89.80	85.70	70.50	67.30
$\eta = 0.00$		78.10	78.10	73.60	73.60		82.90	81.50	**72.70**	**72.70**
$\eta = 0.10$		79.80	79.80	73.90	73.90		82.80	81.40	71.70	71.70
$\eta = 0.25$	Scenario 2	83.00	83.00	75.50	75.50	Scenario 5	85.00	81.80	71.00	71.00
$\eta = 0.40$		**83.50**	**83.50**	**76.20**	**76.20**		**85.70**	**82.50**	70.60	70.60
$\eta = 0.00$		99.10	**99.00**	**97.00**	**97.00**		**98.80**	**96.50**	**93.10**	**92.70**
$\eta = 0.10$		**99.30**	98.60	95.10	95.10		98.60	95.70	89.90	89.60
$\eta = 0.25$	Scenario 3	97.10	96.40	91.90	91.90	Scenario 6	96.20	92.10	87.00	86.20
$\eta = 0.40$		95.90	95.10	89.50	89.50		94.00	89.50	81.90	81.00

combinations. Also, since the scenarios showed in Table 2.5 where generated with a logistic model, we selected the scenarios to observe how robust is our proposal in scenarios generated with other models, such us the Clayton Copula, and scenarios that are arbitrarily generated. Moreover, since we are using the same set of true MTDs as [6, 7], we use these methods to make a performance comparison in terms of percent of correct MTD selection.

In Tables 2.9 and 2.10, in the supplementary material, we present operating characteristics in term of safety and efficiency for each of the 6 proposed scenarios. In general, we observe that the design behaves in a similar way as with the scenarios presented along this manuscript. In terms of safety, toxicity attributions reduce the average percent of toxicities and the percent of trials with toxicity rates greater than $\theta + 0.05$ and $\theta + 0.10$. In terms of efficiency, we only observe a positive effect in scenarios with a relatively flat dose-toxicity surface. In terms of performance comparison, our proposed method is competitive with other standard designs for drug combinations such us [6, 7], and achieves better percent of correct MTD recommendation in 4 out of the 6 used scenarios.

Another issue that is relevant to the methodology we present in this manuscript is the errors in the attribution of toxicities by the treating investigators. Our design does not include a parameter to control the uncertainty around the decision made by the investigator when attributing the the DLT, which could be an extension of this work. However, in the supplementary material, in order to assess the impact of these kind of errors, we present simulation from 3 scenarios taken also from [6, 7] where we introduce 10% and 50% of errors in the attribution of DLTs, and compare it to the case where we correctly attributes

Fig. 2.7 Probability of DLT surfaces of the 6 scenarios from Table 2.5.



100% of the DLTs. In Tables 2.11 and 2.12, we present the simulated results in terms of safety and efficiency. Overall we do not observe any major difference when incorrectly attributing 10% and 50% of the DLTs with respect to correctly attributing 100% of the DLTs.

2.6 Conclusions

In this paper we proposed a Bayesian adaptive design for cancer phase I clinical trials using drug combinations with continuous dose levels and attributable DLT in a fraction of patients. A copula-type model was used to describe the relationship between dose combinations and probability of DLT. The trial design proceeds by treating cohorts of two patients, each patient with a different dose combination estimated using univariate CRM for a better exploration of the space of doses. Treating cohorts of two patients will allow trial conduct to be completed in a reasonable amount of time. Although the two patients in a cohort are allocated to different dose combinations, a patient in the current cohort can be treated at a dose (x,y) if and only if a patient in the previous cohort was treated at a dose on the same horizontal or vertical line within our dose range, that is was treated with either dose x or dose y. The use of continuous dose levels is not uncommon in early phase trials, particularly when the drugs are given as infusions intravenously. For instance, a drug combination trial of cabazitaxel and cisplatin delivered intravenously was recently designed for advanced prostate cancer patients where the dose levels are continuous and the protocol was approved by the scientific review at Cedars-Sinai. For ethical reasons, we further imposed dose escalation restrictions for one of the drugs when a DLT is attributable to that drug.

Table 2.5 Toxicity scenarios for the model misspecification evaluation. In bold the doses considered as true MTD set.

Dose level	1	2	3	4	1	2	3	4
	Scenario 1				Scenario 4			
4	**0.28**	0.41	0.55	0.68	0.04	0.09	0.17	**0.32**
3	**0.25**	**0.35**	0.48	0.60	0.03	0.06	0.12	**0.23**
2	**0.22**	**0.30**	**0.40**	0.51	0.02	0.05	0.09	0.16
1	0.19	**0.26**	**0.34**	0.43	0.02	0.03	0.06	0.11
	Scenario 2				Scenario 5			
4	0.17	**0.29**	0.45	0.62	0.12	**0.26**	0.48	0.71
3	0.14	**0.23**	**0.35**	0.50	0.09	0.19	**0.36**	0.57
2	0.12	0.18	**0.27**	**0.38**	0.07	0.14	**0.26**	0.43
1	0.09	0.14	0.19	**0.27**	0.05	0.10	0.18	**0.30**
	Scenario 3				Scenario 6			
4	**0.37**	0.72	0.92	0.98	0.78	0.94	0.99	1.00
3	**0.26**	0.59	0.85	0.96	0.68	0.90	0.97	0.99
2	0.18	0.44	0.74	0.91	0.57	0.83	0.94	0.98
1	0.12	**0.30**	0.59	0.82	0.45	0.73	0.90	0.97

We studied the operating characteristics of the design under various scenarios for the true location of the MTD curve. In general, we observed that the trial is safe and as the proportion of attributed toxicities increases, the average proportion of toxicities decreases when we attribute toxicities. To assess the efficiency when estimating the MTD curve, we employed the pointwise average bias and average percent selection. In general the method is efficient although the results varied depending on the proportion of attributed toxicities. Note that the operating characteristics were evaluated under vague prior distributions of the model parameters and no toxicity profiles of single agent trials were used *a priori*. We also showed how the method can be adapted to the setting of discrete dose combinations.

We also performed a model misspecification evaluation in scenarios with different dose-toxicity surfaces. We only observed a positive effect in terms of percent of correct MTD recommendation in scenarios with flat surfaces. In scenarios with non-flat dose-toxicity surfaces we observed a decline in performance of percent selection consistent with the findings by [10] when working with copula regression models. We also observed a positive

Table 2.6 Operating characteristics summarizing trial safety for model misspecification in $m = 1000$ simulated trials.

		Average % of toxicities	% of trials with toxicity rate $> \theta + 0.05$	% of trials with toxicity rate $> \theta + 0.10$
Scenario 1	$\eta = 0.00$	32.99	22.90	4.20
	$\eta = 0.10$	32.19	18.50	2.90
	$\eta = 0.25$	31.43	15.80	2.60
	$\eta = 0.40$	30.58	12.90	2.50
Scenario 2	$\eta = 0.00$	29.85	6.60	0.20
	$\eta = 0.10$	29.14	4.10	0.10
	$\eta = 0.25$	28.20	3.10	0.30
	$\eta = 0.40$	27.90	2.30	0.00
Scenario 3	$\eta = 0.00$	36.53	40.70	16.40
	$\eta = 0.10$	35.13	33.90	12.50
	$\eta = 0.25$	33.94	28.60	11.00
	$\eta = 0.40$	32.94	23.40	9.50
Scenario 4	$\eta = 0.00$	22.43	0.00	0.00
	$\eta = 0.10$	21.83	0.00	0.00
	$\eta = 0.25$	21.39	0.00	0.00
	$\eta = 0.40$	20.87	0.00	0.00
Scenario 5	$\eta = 0.00$	30.43	6.60	0.30
	$\eta = 0.10$	29.48	3.30	0.10
	$\eta = 0.25$	28.60	3.60	0.00
	$\eta = 0.40$	27.60	2.30	0.00

effect in scenarios where the lowest dose combination has an excessively high probability of DLT. In this case, toxicity attributions improves the percent of times the trial was stopped due to safety. In all cases, safety of the trial is not compromised by accounting for a partial toxicity attribution. Clearly, there is a trade-off when increasing the fraction of DLT attribution to one or more drugs. The design is more conservative in future escalations, lowering the in-trial DLT percentages and reducing how quickly the MTD contour is reached, by favoring experimentation over recommendation.

Our design is practically useful when the two drugs do not have many overlapping toxicities, see e.g. [34] for some examples of drug combination trials with these characteristics. In cases where we expect a high percent of overlapping DLTs, designs that do not distinguish between drug attribution listed in the introduction may be more appropriate. Our method relies on clinical judgment regarding DLT attribution. In many phase I trials, such decisions are subject to error classifications and a possible extension is to introduce a parameter to

Table 2.7 Percent of times that at least 25%, 50%, 75% or 100% of recommended MTDs belong to the true MTD set in $m = 1000$ simulated trials under model misspecification. In scenario 6 we show the percent of times the trial is stopped due to safety reasons.

		% of correct MTD recommendation for $\theta \pm 0.10$							
		$\geq 25\%$	$\geq 50\%$	$\geq 75\%$	100%		$\geq 25\%$	$\geq 50\%$	$\geq 75\%$	100%
$\eta = 0.00$	Scenario 1	82.90	75.60	55.40	55.40	Scenario 4	**98.80**	**98.80**	**87.80**	**87.80**
$\eta = 0.10$		82.70	72.70	57.30	57.30		97.20	97.20	85.70	85.70
$\eta = 0.25$		**83.30**	**75.80**	**60.40**	**60.40**		95.70	95.70	82.00	82.00
$\eta = 0.40$		80.60	73.10	57.60	57.60		95.20	95.20	76.20	76.20
$\eta = 0.00$	Scenario 2	74.70	71.00	**58.20**	**45.70**	Scenario 5	**75.40**	**69.10**	**20.50**	**20.50**
$\eta = 0.10$		77.00	73.50	53.60	44.60		71.80	62.80	20.40	20.40
$\eta = 0.25$		**79.60**	**75.00**	50.00	41.20		70.70	59.40	18.90	18.90
$\eta = 0.40$		77.30	73.10	47.90	37.80		71.20	60.50	16.70	16.70
$\eta = 0.00$	Scenario 3	**76.90**	**65.30**	**23.30**	**23.30**	Scenario 6	83.60			
$\eta = 0.10$		72.50	61.80	21.90	21.90		82.90			
$\eta = 0.25$		66.40	57.30	18.60	18.60		84.80			
$\eta = 0.40$		66.10	54.70	15.70	15.70		**87.20**			

account for errors in toxicity attribution as in [32] for single agent trials. We also plan to study the performance of this design using other link functions under different copula models, and extend this method to early phase cancer trials with late onset toxicity and by accounting for patient's baseline characteristic by extending the approaches in [14, 35] to the drug combination setting.

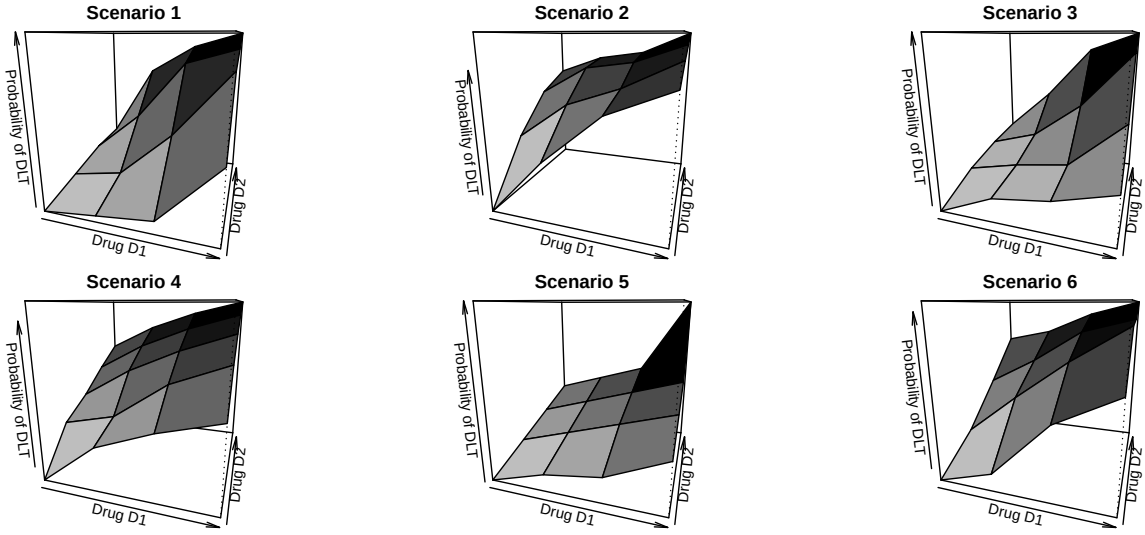
2.7 Supplementary material

In this supplementary section we present additional simulated results that support the conclusions stated in the main document of this article. More precisely, we present two type of simulations to show the robustness of the proposed methodology.

First, we employ six scenarios taken from [6] and [7] where, with respect to the scenarios used in section 5 in the main document, the set of true maximum tolerated doses (MTDs) is smaller (i.e., a more restricted set of MTDs) and also where the difference in terms of probability of toxicity between dose combinations is larger. Moreover, we use these six scenarios to make comparisons in terms of percent of MTD selection with respect to the methodology proposed by [6] and [7].

Second, we employ three scenarios taken from [6], where we introduce 10% and 50% of error in the attribution assessed by the treating investigator. We choose scenarios where

Fig. 2.8 Probability of DLT surfaces of the 3 scenarios from Table 2.8.



toxicity attribution has a positive effect and evaluate the impact of these errors in terms of safety and percent of MTD recommendation.

In Table 2.8, we present the mentioned six scenarios taken from [6] and [7] where we observe a restricted set of MTDs that is also placed in different locations in the space of dose combinations, and also with different probability of toxicity targets. In Figure 2.8 we present the dose-toxicity surface where we observe scenarios with a more flat surface and scenarios with steeper surface.

In Table 2.9 we present operating characteristics in term of safety for each scenario. In general, we observe that toxicity attributions reduce the average percent of toxicities and the percent of trials with toxicity rates greater than $\theta + 0.05$ and $\theta + 0.10$.

In Table 2.10 we the percent of times at least one recommended MTD belongs to the true MTD set. The reason why in this assessment we do not present the percent of times that at least 25%, 50%, 75% or 100% of the recommended MTDs belong to the true MTD set is because this is the only way we can make a fair comparison with other existing methodology since we are able to recommend multiple MTDs at the end of a trial and, for example, [6] and [7], only recommend one MTD at then end of a trial. Nevertheless, we observe a similar behavior as in Tables 4 and 7 in the main document, since there is an improvement in terms of percent of correct MTD selection when the dose-toxicity surface is relatively flat. Compared to [6] and [7], our proposed method is competitive and achieves better performance when $\eta = 0$ in four out of the six employed scenarios.

Table 2.8 Toxicity scenarios for the model misspecification evaluation. In bold the doses considered as true MTD set.

Dose level	1	2	3	4	1	2	3	4	5
	Scenario 1				Scenario 4				
4	**0.30**	0.50	0.55	0.60	0.49	0.58	0.68	0.75	0.81
3	0.12	**0.30**	0.50	0.55	0.48	0.59	0.68	0.75	0.81
2	0.10	0.15	**0.30**	0.45	**0.40**	0.45	0.59	0.67	0.74
1	0.08	0.12	0.16	0.18	0.24	**0.40**	0.47	0.56	0.64
	Scenario 2				Scenario 5				
4	0.48	0.52	0.55	0.58	0.16	0.18	0.20	**0.30**	
3	0.42	0.45	0.50	0.52	0.13	0.16	0.18	0.20	
2	**0.30**	0.40	0.48	0.50	0.12	0.14	0.16	0.18	
1	0.15	**0.30**	0.40	0.45	0.10	0.12	0.14	0.16	
	Scenario 3				Scenario 6				
4	0.20	**0.30**	0.45	0.50	0.50	0.66	0.67	0.73	
3	0.16	0.18	**0.30**	0.45	**0.40**	0.54	0.62	0.68	
2	0.14	0.16	0.20	**0.30**	0.21	**0.40**	0.50	0.60	
1	0.08	0.13	0.16	0.18	0.15	0.25	**0.40**	0.56	

Now we present the simulations related to the errors in the attribution assessed by the treatment investigator. In Tables 2.11 and 2.12 we show the operating characteristing of the design in terms of safety and percent of correct MTD recommendation respectively with 0%, 10% and 50% of attribution misspecification.

In Table 2.11, taking as a reference the case where we correctly attribute all the toxicities, we do not observe any big difference in terms of average percent of toxicities and percent of trials with toxicity rate greater than $\theta + 0.10$, regardless the percentage of attribution misspecification. The only case where we observe a slight worsening is in terms of percent of trials with toxicity rate greater than $\theta + 0.05$.

In Table 2.12, taking again as a reference the case where we correctly attribute all the toxicities, we observe that, when we request that at least 25%, 50%, 75% or 100% of the recommended MTDs belong to the true MTD set, there are no big difference in any of the scenarios, regardless the percent of attribution misspecification, and any difference we may observe are due to the random nature of the simulations.

Table 2.9 Operating characteristics summarizing trial safety for model misspecification in $m = 1000$ simulated trials.

		Average % of toxicities	% of trials with toxicity rate $> \theta + 0.05$	% of trials with toxicity rate $> \theta + 0.10$
$\eta = 0.00$	Scenario 1	31.59	11.60	0.40
$\eta = 0.10$		30.37	7.20	0.30
$\eta = 0.25$		29.42	4.80	0.30
$\eta = 0.40$		28.40	3.90	0.10
$\eta = 0.00$	Scenario 2	36.48	48.59	17.80
$\eta = 0.10$		35.78	42.20	15.00
$\eta = 0.25$		34.56	36.20	9.80
$\eta = 0.40$		33.76	30.00	9.70
$\eta = 0.00$	Scenario 3	28.93	3.60	0.00
$\eta = 0.10$		28.41	2.90	0.10
$\eta = 0.25$		27.32	1.70	0.10
$\eta = 0.40$		26.71	1.30	0.00
$\eta = 0.00$	Scenario 4	49.48	67.30	33.60
$\eta = 0.10$		48.11	59.10	27.50
$\eta = 0.25$		47.31	53.30	25.60
$\eta = 0.40$		46.20	45.60	22.00
$\eta = 0.00$	Scenario 5	21.49	0.00	0.00
$\eta = 0.10$		21.07	0.00	0.00
$\eta = 0.25$		20.68	0.00	0.00
$\eta = 0.40$		20.32	0.00	0.00
$\eta = 0.00$	Scenario 6	45.32	38.60	12.00
$\eta = 0.10$		43.99	30.90	8.10
$\eta = 0.25$		42.57	23.50	6.70
$\eta = 0.40$		41.11	18.00	4.50

Table 2.10 Percent of times that at least one of the recommended MTDs belong to the true MTD set in $m = 1000$ simulated trials.

		% of correct MTD selection	
		Partial toxicity attribution	*Yin and Yuan (2009a) †Yin and Yuan (2009b)
$\eta = 0.00$	Scenario 1	**67.50**	53.00*
$\eta = 0.10$		**67.50**	
$\eta = 0.25$		66.00	
$\eta = 0.40$		62.10	
$\eta = 0.00$	Scenario 2	46.10	49.80*
$\eta = 0.10$		**50.80**	
$\eta = 0.25$		50.50	
$\eta = 0.40$		48.50	
$\eta = 0.00$	Scenario 3	**53.50**	50.70*
$\eta = 0.10$		51.30	
$\eta = 0.25$		52.50	
$\eta = 0.40$		51.90	
$\eta = 0.00$	Scenario 4	35.10	44.00†
$\eta = 0.10$		**35.70**	
$\eta = 0.25$		34.40	
$\eta = 0.40$		33.00	
$\eta = 0.00$	Scenario 5	64.10	60.50*
$\eta = 0.10$		**72.80**	
$\eta = 0.25$		67.00	
$\eta = 0.40$		60.40	
$\eta = 0.00$	Scenario 6	59.10	57.50†
$\eta = 0.10$		60.00	
$\eta = 0.25$		**61.20**	
$\eta = 0.40$		57.40	

Table 2.11 Operating characteristics summarizing trial safety in $m = 1000$ simulated trials under attribution misspecification.

Table	Scenario	% of attribution misspecification	Average % of toxicities	% of trials with toxicity rate $> \theta + 0.05$	% of trials with toxicity rate $> \theta + 0.10$
2.5	1	0	31.43	15.80	2.60
		10	31.67	18.10	2.70
		50	32.05	18.40	2.40
2.8	2	0	35.78	42.20	15.00
		10	35.59	42.40	13.80
		50	36.09	45.70	16.40
2.8	5	0	21.07	0.00	0.00
		10	21.17	0.00	0.00
		50	21.28	0.00	0.00

Table 2.12 Percent of times that at least 25%, 50%, 75% or 100% of recommended MTDs belong to the true MTD set in $m = 1000$ simulated trials under attribution misspecification.

			% of MTD recommendation			
Table	Scenario	% of attribution misspecification	$\geq 25\%$	$\geq 50\%$	$\geq 75\%$	100%
2.5	1	0	83.30	75.80	60.40	60.40
		10	83.60	74.80	59.00	59.00
		50	83.99	75.60	57.90	57.90
2.8	2	0	50.80	50.80	34.80	34.80
		10	49.70	49.70	34.90	34.90
		50	48.60	48.60	31.60	31.60
2.8	5	0	72.80	72.80	72.80	72.80
		10	74.40	74.40	74.40	74.40
		50	74.20	74.20	74.20	74.20

Chapter 3

A Bayesian two-stage adaptive design for cancer phase I/II trials with drug combinations

3.1 Introduction

In cancer phase I/II clinical trials, the main goal is to identify a safe dose that maximizes the treatment efficacy. In single-agent settings with binary or time to event endpoints where efficacy is observed relatively fast (e.g. one or two cycles of therapy), one-stage sequential designs where the joint probability of toxicity and efficacy is sequentially updated after each cohort of patients are usually employed (see e.g. [31, 36–41] for binary endpoints, and [42] for time to event endpoints). This methodology has been extended to accommodate combination of drugs of any kind (see e.g. [43, 44] for binary endpoints and [42] for time to event endpoints), and proceed in a similar fashion as the methods referenced for single-agent.

In cases where efficacy is not ascertained in a short period of time, it is frequent to employ two-stage designs where, a maximum tolerated dose (MTD) set is first selected, and then tested for efficacy in a second stage with possibly a different population of patients than the one used in the first stage. This approach has been discussed by [45–47]. For drug combination trials, methodology for these type of two-stage designs have been proposed for binary efficacy endpoints (see e.g. [48, 49]).

One characteristic that most of these methods have in common is that they only recommend a single MTD either at the end of the phase I trial, or at the end of the first stage in a

phase I/II trial. However, even if the recommended dose that will be tested for efficacy is indeed a valid MTD, there could be another MTD with higher efficacy, making the MTD recommended in the first place non-optimal.

In this article, we present a two-stage design for drug combination trials

In this article, we extend the work from [49] by proposing a two-stage design for drug combinations trials with time to event efficacy endpoint in the second stage and continuous dose levels when treatment efficacy is evaluated after three or more cycles of therapy. In the first stage, the dose finding method proposed by [13] is used to estimate the MTD curve. In the second stage, a Bayesian adaptive design that starts allocating a first cohort of patients to dose combinations equally spaced along the MTD curve, and then allocates subsequent cohorts of patients to dose combinations likely to have high posterior median TTP using adaptive randomization. To allow for different shapes in the median TTP curve, we employ a flexible family of cubic splines to model the dose - median TTP relationship. Adaptive randomization is sequentially used after a pre-defined time period to minimize the number of patients allocated at sub-therapeutic dose levels. At the end of the trial, the dose combination within the MTD with highest *a posteriori* median TTP is selected and recommended for further phase II or III studies.

The manuscript is organized as follows. In section 2, we review the first stage of the proposed phase I/II trial previously described in [13, 49]. In section 3, we describe the second stage of the design. In section 4, we illustrate the methodology with the phase I/II drug combination trial of cisplatin and cabazitaxel in patients with prostate cancer with visceral metastasis where time to progression is a secondary endpoint. The goal in this trial is to find a tolerable dose combination with highest TTP median. A discussion of the approach and final remarks are included in Section 5.

3.2 Phase I/II Trial: Stage 1

3.2.1 Model

Following [13], consider the generic form of a dose-toxicity model

$$P(T = 1|x,y) = F(\eta_0 + \eta_1 x + \eta_2 y + \eta_3 xy), \tag{3.1}$$

where $T = 1$ represents an observed DLT at the dose combination (x, y), $T = 0$ otherwise, $x \in [X_{\min}, X_{\max}]$ is the dose level of agent A, $y \in [Y_{\min}, Y_{\max}]$ is the dose level of agent B and $F(.)$ is a known cumulative distribution function. We assume that the dose combinations are continuous and standardized to be in the interval [0,1], the interaction parameter $\eta_3 > 0$, and $\eta_1, \eta_2 > 0$ in order to guarantee that the probability of DLT increases with the dose of any agent when the other one is held constant.

The MTD is defined as any dose combination (x^*, y^*) such that

$$P(T = 1|x^*, y^*) = \theta. \tag{3.2}$$

As described in [13], we reparameterize equation (3.1) as follows. Let ρ_{10}, the probability of DLT when the levels of drugs A and B are 1 and 0, respectively, ρ_{01}, the probability of DLT when the levels of drugs A and B are 0 and 1, respectively, ρ_{00}, the probability of DLT when the levels of drugs A and B are both 0. Hence, it is possible to show that MTD takes the form

$$\begin{aligned} C = \Big\{(x^*, y^*) : y^* = \Big[(F^{-1}(\theta) - F^{-1}(\rho_{00})) - (F^{-1}(\rho_{10}) - F^{-1}(\rho_{00}))x^*\Big] \\ \div \Big[(F^{-1}(\rho_{01}) - F^{-1}(\rho_{00}))\eta_3 x^*\Big]\Big\}. \end{aligned} \tag{3.3}$$

We assume that ρ_{10}, ρ_{01} and η_3 are independent *a priori* with $\rho_{01} \sim \text{beta}(a_1, b_1)$, $\rho_{10} \sim \text{beta}(a_2, b_2)$, and conditional on (ρ_{01}, ρ_{10}), $\rho_{00}/\min(\rho_{01}, \rho_{10}) \sim \text{beta}(a_3, b_3)$. The prior distribution on the interaction parameter η_3 is a gamma distribution with mean a/b and variance a/b^2. Let $D_n = \{(x_i, y_i, T_i)\}$ be the data gathered after enrolling n patients. The posterior distribution of the model parameters is

$$\begin{aligned} \pi(\rho_{00}, \rho_{10}, \rho_{01}, \eta_3) &\propto \prod_{i=1}^{n} G((\rho_{00}, \rho_{10}, \rho_{01}, \eta_3; x_i, y_i))^{T_i} \\ &\times (1 - G(\rho_{00}, \rho_{10}, \rho_{01}, \eta_3; x_i, y_i))^{1-T_i} \\ &\times \pi(\rho_{01})\pi(\rho_{10})\pi(\rho_{00}|\rho_{01}, \rho_{10})\pi(\eta), \end{aligned} \tag{3.4}$$

where

$$G(\rho_{00},\rho_{10},\rho_{01},\eta_3;x_i,y_i) = F(F^{-1}(\rho_{00}) + (F^{-1}(\rho_{10}) - F^{-1}(\rho_{00}))x_i + (F^{-1}(\rho_{01}) - F^{-1}(\rho_{00}))y_i + \eta_3 x_i y_i). \tag{3.5}$$

Note that the operating characteristics of this stage are evaluated using informative prior distributions (see [49]).

3.2.2 Trial Design

Dose escalation / de-escalation proceeds using the same algorithm described in [13]. It is based on escalation with overdose control (EWOC) where, after each cohort of enrolled patients, the posterior probability of overdosing the next cohort of patients is bounded by a feasibility bound α, see e.g. [11, 50–52]. In a cohort with two patients, the first one would receive a new dose of agent A given that the dose y of agent B that was previously assigned. The other patient would receive a new dose of agent B given that dose x of agent A was previously assigned. Using EWOC, these new doses are at the α-th percentile of the conditional posterior distribution of the MTDs. The algorithm continues until the maximum sample size is reached or until the trial is stopped for safety. A detailed description of this algorithm can be found in [13]. At the end of the trial, the MTD curve is estimated as

$$C_{\text{est}} = \left\{(x^*,y^*) : y^* = \left[(F^{-1}(\theta) - F^{-1}(\widehat{\rho}_{00})) - (F^{-1}(\widehat{\rho}_{10}) - F^{-1}(\widehat{\rho}_{00}))x^*\right] \div \left[(F^{-1}(\widehat{\rho}_{01}) - F^{-1}(\widehat{\rho}_{00}))\widehat{\eta}_3 x^*\right]\right\}, \tag{3.6}$$

where $\widehat{\rho}_{00},\widehat{\rho}_{10},\widehat{\rho}_{01},\widehat{\eta}$ are the posterior medians given the data D_n. This method has been extensively studied by [13] and hence we only present the operating characteristics in the context of the CisCab trial described in section 3.4.

3.3 Phase I/II Trial: Stage 2

3.3.1 Model

Let x be a dose of drug A such that $(x,y) \in C_{\text{est}}$. Also, assume that x is standardized to in the interval [0,1]. We model the time to progression as a Weibull distribution with probability density function

$$f(t;x) = \frac{k}{\lambda(x;\psi)} \left(\frac{t}{\lambda(x;\psi)} \right)^{k-1} \exp\left(-\frac{t}{\lambda(x;\psi)} \right)^{k}. \tag{3.7}$$

The median TTP is

$$\text{Med}(x) = \lambda(x;\psi)(\log 2)^{\frac{1}{k}}. \tag{3.8}$$

A flexible way of modeling the median TTP along the MTD curve is through the use of the cubic spline function

$$\lambda(x;\psi) = \exp\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{j=3}^{k} \beta_j (x-\kappa_j)_+^3 \right), \tag{3.9}$$

where $\psi = (\beta, \kappa)$, with $\beta = (\beta_0, \ldots, \beta_k)$ and $\kappa = (\kappa_3, \ldots, \kappa_k)$, being $\kappa_3 = 0$. Let $D_m = \{(x_i, t_i, \delta_i), i = 1 \ldots, m\}$ be the data after enrolling m patients in the trial where t represents the TTP or last follow-up, and δ the censoring status, and let $\pi(\psi, k)$ be the joint prior density on the parameter vector ψ and k. The posterior distribution is

$$\begin{aligned} \pi(\psi, k|D_m) \propto \pi(\psi,k) \prod_{i=1}^{m} \left[\frac{k}{\lambda(x_i;\psi)} \left(\frac{t_i}{\lambda(x_i;\psi)} \right)^{k-1} \right]^{\delta_i} \\ \times \exp\left(-\frac{t_i}{\lambda(x_i;\psi)} \right)^{k}. \end{aligned} \tag{3.10}$$

Let Med_x be the median TTP at dose combination x and let Med_0 be the median TTP of the standard of care treatment. We propose an adaptive design in order to test the hypothesis

$$\begin{aligned} &\mathrm{H}_0 : \mathrm{Med}_x \leq \mathrm{Med}_0 \text{ for all } x \quad \text{vs.} \\ &\mathrm{H}_1 : \mathrm{Med}_x > \mathrm{Med}_0 \text{ for some } x. \end{aligned} \tag{3.11}$$

3.3.2 Trial Design

i We first treat n_1 patients at dose combinations $x_1, \ldots, x_{n_1}$, which are equally spaced along the estimated MTD curve C_{est}.

ii Obtain Bayes estimates of $\hat{\psi}$ and $\hat{k}$, of ψ and k given the data D_{n_1} using equation (3.10). Note that prior to obtaining the Bayes estimates, patients that have not progressed are censored.

iii Generate n_2 dose combinations from the standardized density $\hat{\mathrm{Med}}(x) = \lambda(x;\hat{\psi})(\log 2)^{\frac{1}{\hat{k}}}$ and assign them to the next n_2 patients.

iv Repeat steps (ii) and (iii) until a total of n patients have been enrolled to the trial subject to pre-specified stopping rules.

Decision Rule: At the end of the trial, we reject the null hypothesis if $\mathrm{Max}_x\{P(\mathrm{Med}(x;\psi_i) > \mathrm{Med}_0|D_{n,i})\} > \delta_u$, where δ_u is a design parameter.

Stopping Rule (Futility): For ethical reasons and to avoid treating patient at sub-therapeutic dose levels, we will stop the trial for futility if there is strong evidence that none of the dose combinations are promising, i.e., $\mathrm{Max}_x\{P(\mathrm{Med}(x;\psi_i) > \mathrm{Med}_0|D_{n,i})\} < \delta_0$, where δ_0 is a design parameter.

Stopping Rule (Efficacy): For ethical reasons, if the investigator considers there is enough evidence in favor of one or more dose combinations being tested, and no further patients need to be enrolled, the trial can be terminated if $\mathrm{Max}_x\{P(\mathrm{Med}(x;\psi_i) > \mathrm{Med}_0|D_{n,i})\} > \delta_1$, where $\delta_1 \geq \delta_u$ is a study parameter and the dose combination
$x^{\text{opt}} = \arg\max_v\{P(\mathrm{Med}(v;\psi_i) > \mathrm{Med}_0|D_{n,i})\}$ is selected for further randomized phase II or phase III clinical trials.

Design Operating Characteristics

We assess the operating characteristics of the proposed design by assuming that $\lambda(x;\psi)$ is a cubic spline with two knots placed between 0 and 1. This class of modeling is very flexible

and is able to adapt to scenarios where the median TTP is either constant or skewed toward one of the edges. Vague priors are placed on the model parameters by assuming $\beta \sim N(\mu, \sigma^2 I_6)$, where $\mu = \{0,0,0,0,0,0\}$ and $\sigma^2 = 100$, and $(\kappa_4, \kappa_5) \sim \text{Unif}\{(u,v) : 0 \leq u < v \leq 1\}$. Note that the parameters of the prior distribution of β are always the same regardless the value of Med_0.

For each scenario favoring the alternative hypothesis, we estimate the Bayesian power, which is defined as

$$\text{Power} \approx \frac{1}{M} \sum_{i=1}^{M} I[\text{Max}_x\{P(\text{Med}(x;\psi_i) > \text{Med}_0 | D_{n,i})\} > \delta_u], \tag{3.12}$$

where

$$P(\text{Med}(x;\psi_i) > \text{Med}_0 | D_{n,i}) \approx \frac{1}{L} \sum_{j=1}^{L} I\left[\text{Med}(x;\psi_{i,j}) > \text{Med}_0\right] \tag{3.13}$$

and $\psi_{i,j}$ is the j-th MCMC sample for the i-th trial.

For scenarios favoring the null hypothesis, (3.12) is the estimated Bayesian type-I error probability. The optimal dose from the i-th trial is defined as

$$x_i^{\text{opt}} = \arg\max_v\{P(\text{Med}(v;\psi_i) > \text{Med}_0 | D_{n,i})\}. \tag{3.14}$$

We also report the estimated TTP median by replacing ψ in (3.8) by the average posterior median across all simulated trials. Last, we also report the mean posterior probability of $\text{Med}_x > \text{Med}_0$ for any dose combination x.

3.4 Application to the CisCab Phase I/II Trial

We illustrate the methods proposed in sections 3.2 and 3.3 with a phase I/II trial referred as the "CisCab trial" where TTP is a secondary endpoint of the trial. We are motivated by a phase I trial published by [53], that combines cisplatin and cabazitaxel in patients with advanced solid tumors, where the MTD was established at 15/75 mg/m^2. In a first part of this motivating trial, doses were escalated according to a standard "3+3" design and no DLTs were observed at dose combination which was found to be the MTD. During the second part of the trial, 15 additional patient were treated at the MTD and 2 DLTs were observed. In total, 18 patients were treated at the MTD.

Considering these results, there may be other active dose combinations that are tolerable and active in prostate cancer with visceral metastasis. The CisCab trial considers doses that range from 10 to 25 mg/m^2 for cabazitaxel, and from 50 to 100 mg/m^2 for cisplatin, that will be administered intravenously. In a first stage, the CisCab trial will enroll 30 patients in order to obtain the MTD curve. This stage of the design proceeds as explained in section 3.3, with a target probability of DLT $\theta = 0.33$, and a logistic link fuction $F(.)$ in equation (3.1). The starting dose combination for the first cohort of two patients is 15/75 mg/m^2, and DLTs are to be resolved within 1 cycle of treatment (3 weeks). Prior distributions are calibrated such that the prior mean probability of DLT at dose combination 15/75 mg/m^2 equals θ (see [49]). The operating characteristics of this first stage are obtained by simulating 1000 trials replicates following [13].

In Figures 1 and 2 we show the true and estimated MTD curves obtained with equation (3.6) respectively in two different scenarios . In the scenario presented in Figure 1, the true MTD curve passes through the dose combination 15/75 mg/m^2, whereas in the scenario presented in Figure 2, the true MTD curve is significantly above the dose combination 15/75 mg/m^2. In both scenarios the estimated MTD curves are very close to the true MTD curves. These results are supported the pointwise average bias shown in Figures 6 and 7. In these scenarios, the pointwise average bias fluctuates between -0.01 and 0.01, and between -0.05 and 0.1 respectively. In terms of safety, the percent of trials with DLT rate above $\theta + 0.1$ is below 10% in both scenarios with average number of DLTs of 34% and 27%. We also present results regarding percent correct recommendation. These results are shown in Figures 8 and 9 and overall we observe that, in the two proposed scenarios, the percent of correct recommentation is between 70% and 100% in the scenario where the MTD passes through the dose combination 15/75 mg/m^2, and between 50% and 100% in the scenario where the MTD is above the dose combination 15/75 mg/m^2. Note that these results depend on a design parameter p that takes the values 0.05 and 0.1and that states how strict we are when considering a correct recommendation, being $p = 0.1$ less strict than $p = 0.05$. The true parameter values as well as the safety results are shown in Table 3.1.

In the second stage, 30 additional patients are enrolled to identify the dose combinations along the MTD curve from the first stage, that are likely to have high posterior median TTP. The TTP of the standard care of treatment, which is necessary to perform the hypothesis testing procedure, is chosen to be 4 months since this is the radiographic median TTP in a placebo arm in a previous phase III trial. We present simulations based on 4 scenarios supporting the alternative hypothesis and 4 scenarios favoring the null hypothesis. For each scenario favoring the alternative hypothesis, effect sizes of 1.5 and 2 months and accrual rates of 1 and 2 patients per month will be used. In order to correctly assess the operating

characteristics of the design, the 4 scenarios will have the same TTP of the standard treatment of care, the same effect size and the same accrual rate. This way, the only difference between scenarios will be the shape of the TTP median curve, allowing to see the behavior of the design when the optimal dose level is located at different dose levels.

The simulations were carried out using the model and prior distributions presented in sections 3.3.1 and 3.3.2 respectively, with $n_1 = 10$, $n_2 = 5$, $\delta_u = 0.8$ and $\delta_u = 0.9$.

In Figures 3 and 4 we present the 4 simulated scenarios favoring the alternative hypothesis with effect sizes of 1.5 and 2 months respectively, and an accrual rate of 1 patient per month. In Figure 5 we present the same 4 simulated scenarios favoring the null hypothesis with an accrual rate of 1 patient per month. Results for the same 4 simulated scenarios with an accrual rate of 2 patients per month can be found in Figures 10, 11 and 12 in the appendix. In these figures, we present the true median TTP curve, the null medial TTP, as well as the average recommended dose, the estimated median TTP curve and the posterior probability that the median TTP at a dose level x is greater than the null median TTP as measurements of efficiency. Overall, we observe that the design captures the shape of the median TTP curve. However, the estimated median TTP curve is not a very informative measurement of efficiency since we are using adaptive randomization and hence much more patients are allocated in certain dose levels. The efficiency measurement we believe is more interesting is the posterior probability that the median TTP at a dose level x is greater than the TTP of the standard treatment of care. In all figures, we observe that the recommended optimal dose is very close to the true optimal dose regardless the shape of the TTP median, the effect size or the accrual rate.

In Table 3.2, we present the Bayesian power, the probability of the type-I error as well as the probability of type-I + type-II errors for different effect sizes and different accrual rates. With an accrual rate of 1 patient per month, the probability of type-I error remains between 0.104 and 0.227 when $\delta_u = 0.8$ and between 0.235 and 0.308 when $\delta_u = 0.9$. However, with an accrual rate of 2 patients per month, the probability of type-I error is much smaller overall and it remains between 0.035 and 0.107 when $\delta_u = 0.8$ and between 0.008 and 0.048 when $\delta_u = 0.9$.

In terms of power, with effect size of 1.5 months and an accrual rate of 1 patient per month, we observe that the power remains between 0.706 and 0.924 when $\delta_u = 0.8$ and between 0.52 and 0.844 when $\delta_u = 0.9$. If the effect size increases up to 2 months and maintaining the same accrual rate, the power remains between 0.931 and 0.972 when $\delta_u = 0.8$ and between 0.846 and 0.932 when $\delta_u = 0.9$. In contrast, if we fix the accrual rate to 2 patient per month, we observe that overall the power decreases considerately. With an effect size of 1.5 months,

the power remains between 0.522 and 0.824 when $\delta_u = 0.8$ and between 0.338 and 0.674 when $\delta_u = 0.9$. If the effect size increases up to 2 months and maintaining the same accrual rate, the power remains between 0.766 and 0.92 when $\delta_u = 0.8$ and between 0.615 and 0.829 when $\delta_u = 0.9$.

Because it is difficult to find the right balance between power and type-I error, and since it is not unusual to find probabilities of type-I error between 0.15 - 0.2 in phase II trials of these characteristics where we try a large set of doses with a small sample size, we evaluate the sum of the probabilities of type-I error and type-II errors. In general, a design where the sum of these two probabilities is above 0.3 is not advisable. In our proposal, with effect size of 1.5 months and an accrual rate of 1 patient per month, the sum of the probabilities of type-I error and type-II error remains between 0.235 and 0.416 when $\delta_u = 0.8$ and between 0.264 and 0.523 when $\delta_u = 0.9$. If the effect size increases up to 2 months and maintaining the same accrual rate, the sum of the probabilities of type-I error and type-II error remains between 0.15 and 0.256 when $\delta_u = 0.8$ and between 0.123 and 0.198 when $\delta_u = 0.9$. If we fix the accrual rate to 2 patient per month, with an effect size of 1.5 months, the sum of the probabilities of type-I error and type-II error remains between 0.283 and 0.513 when $\delta_u = 0.8$ and between 0.374 and 0.67 when $\delta_u = 0.9$. If the effect size increases up to 2 months and maintaining the same accrual rate, the the sum of the probabilities of type-I error and type-II error remains between 0.17 and 0.287 when $\delta_u = 0.8$ and between 0.219 and 0.403 when $\delta_u = 0.9$.

If we focus one the sum of the probabilities of type-I error and type-II error, we observe that with an effect size of 1.5 months we observe a lot of values above our 0.3 threshold regardless the accrual rate, which is normal since the original design's primary endpoint was not the TTP median and it is not sufficiently powered for this effect size. In contrast, with an effect size of of 2 months, we observe that if $\delta_u = 0.8$ all the values are below our 0.3 threshold regardless the accrual rate and if $\delta_u = 0.9$, only one scenario with an accrual rate of 2 patients per month has a value above the threshold.

In Table 3.3, we present the probability of early stopping and average sample size at the moment of stopping in scenarios favoring the null hypothesis. Overall, we observe that an accrual rate of 2 patients per month produces a slight increase in the probability of early stopping and a decrease between 1 and 2 patients in the average sample size at the moment of stopping with respect to using an accrual rate of 1 patient per month.

Even though it is not listed in the operating characteristics, in Figure 13 we show the dose allocation distribution in the four scenarios with the different effect sizes and accrual rates. In scenario 1, we correctly allocate more than 71% of the patients in doses that are above the

TTP of the standard treatment of care. In scenarios 2, 3 and 4 we correctly allocate more than 65%, 77% and 60% of the patients respectively in dose above the TTP of the standard treatment of care. Note that from these distributions we excluded the first n_1 doses which are automatically allocated in doses equally spaced along the MTD.

We also implement scenarios 1 and 2 when the TTP of the standard treatment of care is higher than 4 months. More precisely we tuned scenarios 1 and 2 to have effect sizes of 2 months but the TTP of the standard treatment of care is now 8 months. We used accrual rates of 2 and 3 patients per month and we observed values of power, type-I error and sum of type-I and type-II errors that are consistent with the values presented in Table 3.2. These results are showed in Table 3.4.

Hence, we conclude that our design has overall good operating characteristics with accrual rates that are considered realistic in practice.

3.5 Conclusions

In this paper we propose a Bayesian adaptive two-stage design for cancer phase I/II trials using drug combinations with continuous dose levels and TTP endpoint. We are motivated by a phase I trial published by [53], that combines cisplatin and cabazitaxel in patients with prostate cancer with visceral metastasis, where the MTD was established at 15/75 mg/m^2. In a first part of this motivating trial, doses were escalated according to a standard "3+3" design and no DLTs were observed at dose combination which was found to be the MTD. During the second part of the trial, 15 additional patient were treated at the MTD and 2 DLTs were observed, and in total 18 patients were treated at the MTD. However, considering these results, there may be other active dose combinations that are tolerable and active in prostate cancer with visceral metastasis.

In the first stage of the design a logistic model is used to model the probability of DLT. The dose escalation algorithm proceeds by using EWOC as described in [13]. At the end of this stage, an estimate of the MTD curve is obtained. In the second stage we model the median TTP along the MTD curve using a weibull model and incorporating a cubic spline through the scale parameter of the model. In this stage of the design a hypothesis test is performed where the null hypothesis states that the median TTP corresponding to all dose levels is below or equal to a TTP of the standard treatment of care. On the other hand, the alternative hypothesis states that the median TTP corresponding to some dose levels is above the TTP of the standard treatment of care. The dose escalation in the second stage

proceeds by first allocating n_1 patients in dose levels equally spaced along the MTD curve. Subsequent patients are allocated in cohorts of n_2 patients in doses with higher posterior probability of having a median TTP greater TTP of the standard treatment of care using adaptive randomization.

Regarding the first stage, we studied the operating characteristics in 2 scenarios. In the one scenario the true MTD curve passes through the dose combination 15/75 mg/m^2, whereas in the other scenario the true MTD curve is significantly above the dose combination 15/75 mg/m^2. We found that this stage of the trial is safe and has good operating characteristics in terms of pointwise bias and percent selection. Note that the operating characteristics of this stage were evaluated using informative prior distributions as commented in section 3.4.

With respect to the second stage, we studied the operating characteristics of the design in 4 scenarios in which the null median TTP is the same and so is the effect size and accrual rate and hence the only difference between them is the median TTP curve shape that places the dose level with highest TTP at a different location in each scenario. These 4 scenarios were implemented with effects sizes of 1.5 and 2 months, and accrual rates of 1 patient and 2 patients per month, which are considerate reasonable in practice. In general, we observed good operating characteristics in terms of optimal dose recommendation and sum of the probabilities of type-I and type-II errors as main measurements of efficiency. Scenarios 1 and 2 were also implemented when the TTP of the standard treatment of care is higher than 4 months. More precisely we tuned scenarios 1 and 2 to have effect sizes of 2 months but the TTP of the standard treatment of care is now 8 months. We used accrual rates of 2 and 3 patients per month and we observed values of power, type-I error and sum of type-I and type-II errors that are consistent with the values presented in Table 3.2.

Note that the operating characteristics in the second stage were evaluated under vague prior distributions of the model parameters and no efficacy profiles of single agent trials we used *a priori*.

A limitation of this methodology is that the uncertainty of the estimated MTD curve in stage I is not taken into account in stage II of the design. This implies that the MTD curve is not updated during the second stage, which is a limitation since patients in stage II may come from a different population with respect to patients in the first stage. As pointed out by [49], an alternative design for this particular paper would account for first-, second- and third-cycle DLT in addition to efficacy outcome at each cycle. Also, the nature of DLT (reversible versus non-reversible) should be taken into account since patients with a reversible DLT are usually treated for that side effect and kept in the trial with dose reduction in subsequent cycles. Hence, for drug combinations with continuous dose levels and three cycles of therapy,

another layer of model complexity would be introduced but such designs are beyond the scope of this paper and are subjects of future research. In addition, we note that a continuous monitoring of the rate of DLT in stage II is also carried out as discussed in [49] so that the trial will stop early if there is evidence of an excessive rate of DLT.

Fig. 3.1 True and estimated MTD curve for scenario 1 in the first stage of the design.

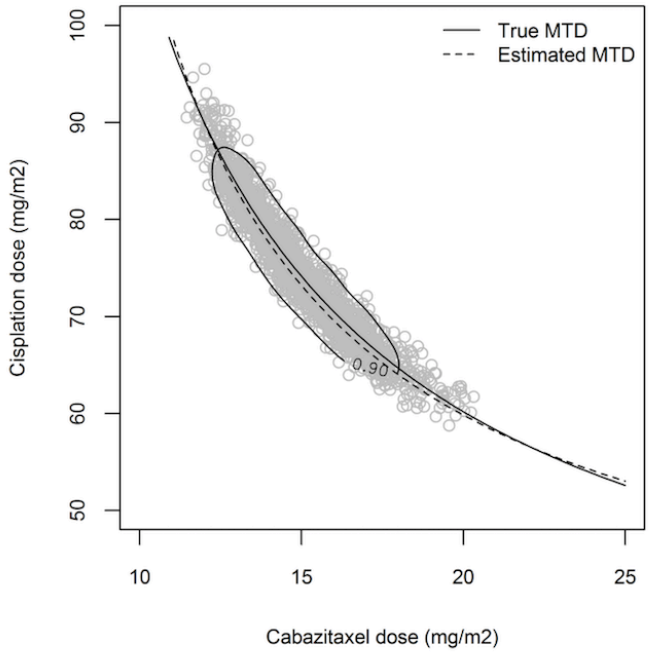


Table 3.1 True parameter values for $\rho_{00}, \rho_{01}, \rho_{10}, \eta$ and θ as well as safety results for the two simulated scenarios of the first stage where EWOC is employed.

	Scenario 1	Scenario 2
ρ_{00}	1e-5	1e-8
ρ_{01}	0.10	0.00005
ρ_{10}	0.10	0.00008
η	20	20
θ	0.33333	0.33333
Average number of DLTs	0.34	0.27
Number of trials with DLT rate $> \theta + 0.1$	7.30	0.00

Fig. 3.2 True and estimated MTD curve for scenario 2 in the first stage of the design.

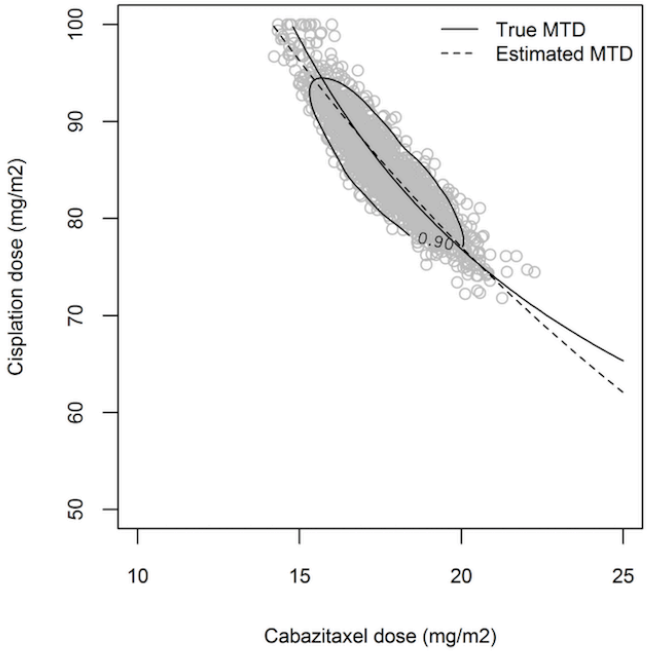


Table 3.2 Bayesian power, type I error probability and type-I + type-II error probability in four scenarios with effect sizes of 1.5 and 2 months, and accrual rates of 1 and 2 patients per month.

		Power (effect size of 1.5 months)		Power (effect size of 2 months)		Probability of type-I error		Probability of type-I + type-II errors (effect size of 1.5 months)		Probability of type-I + type-II errors (effect size of 2 months)	
		δ_u		δ_u		δ_u		δ_u		δ_u	
Scenario	Accrual rate	0.8	0.9	0.8	0.9	0.8	0.9	0.8	0.9	0.8	0.9
1	1	0.924	0.844	0.971	0.927	0.227	0.121	0.303	0.277	0.256	0.194
2		0.706	0.520	0.972	0.920	0.122	0.043	0.416	0.523	0.150	0.123
3		0.904	0.808	0.973	0.932	0.139	0.072	0.235	0.264	0.166	0.140
4		0.796	0.646	0.931	0.846	0.104	0.044	0.308	0.398	0.173	0.198
1	2	0.824	0.674	0.920	0.829	0.107	0.048	0.283	0.374	0.187	0.219
2		0.522	0.338	0.865	0.755	0.035	0.008	0.513	0.670	0.170	0.253
3		0.759	0.598	0.896	0.790	0.068	0.024	0.309	0.426	0.172	0.234
4		0.623	0.445	0.766	0.615	0.053	0.018	0.430	0.573	0.287	0.403

Fig. 3.3 True and estimated TTP medians under four scenarios favoring the alternative hypothesis with effect size of 1.5 months and accrual rate of 1 patient per month.

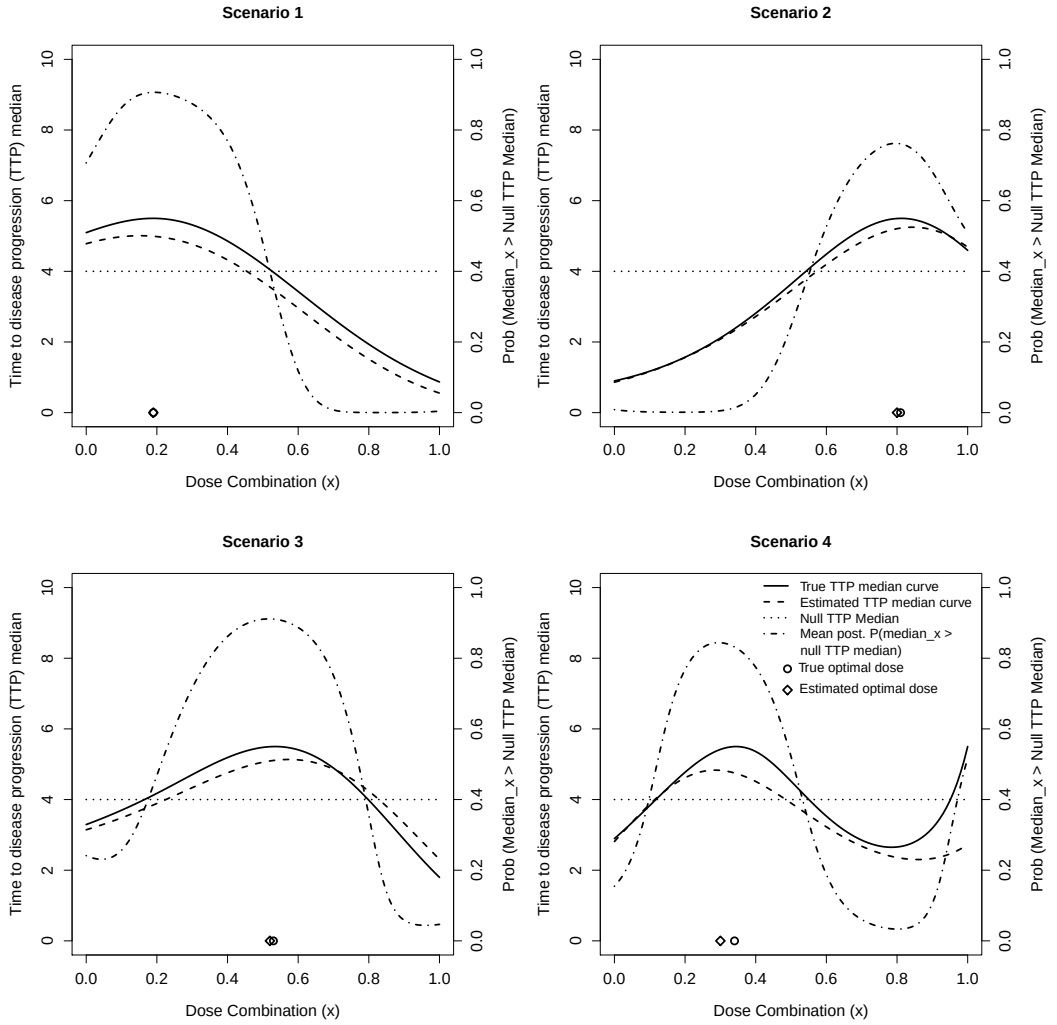


Fig. 3.4 True and estimated TTP medians under four scenarios favoring the alternative hypothesis with effect size of 2 months and accrual rate of 1 patient per month.

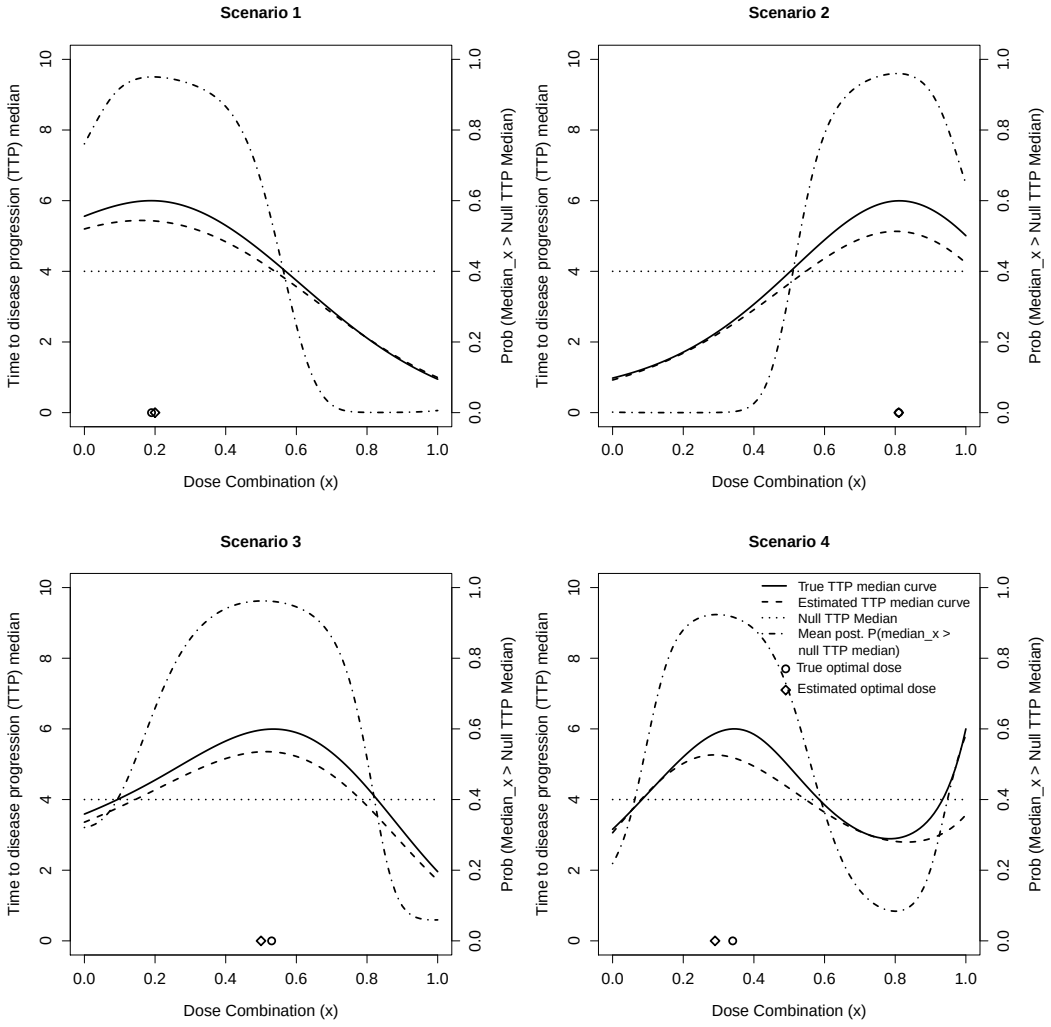


Fig. 3.5 True and estimated TTP medians under four scenarios favoring the null hypothesis with an accrual rate of 1 patient per month.

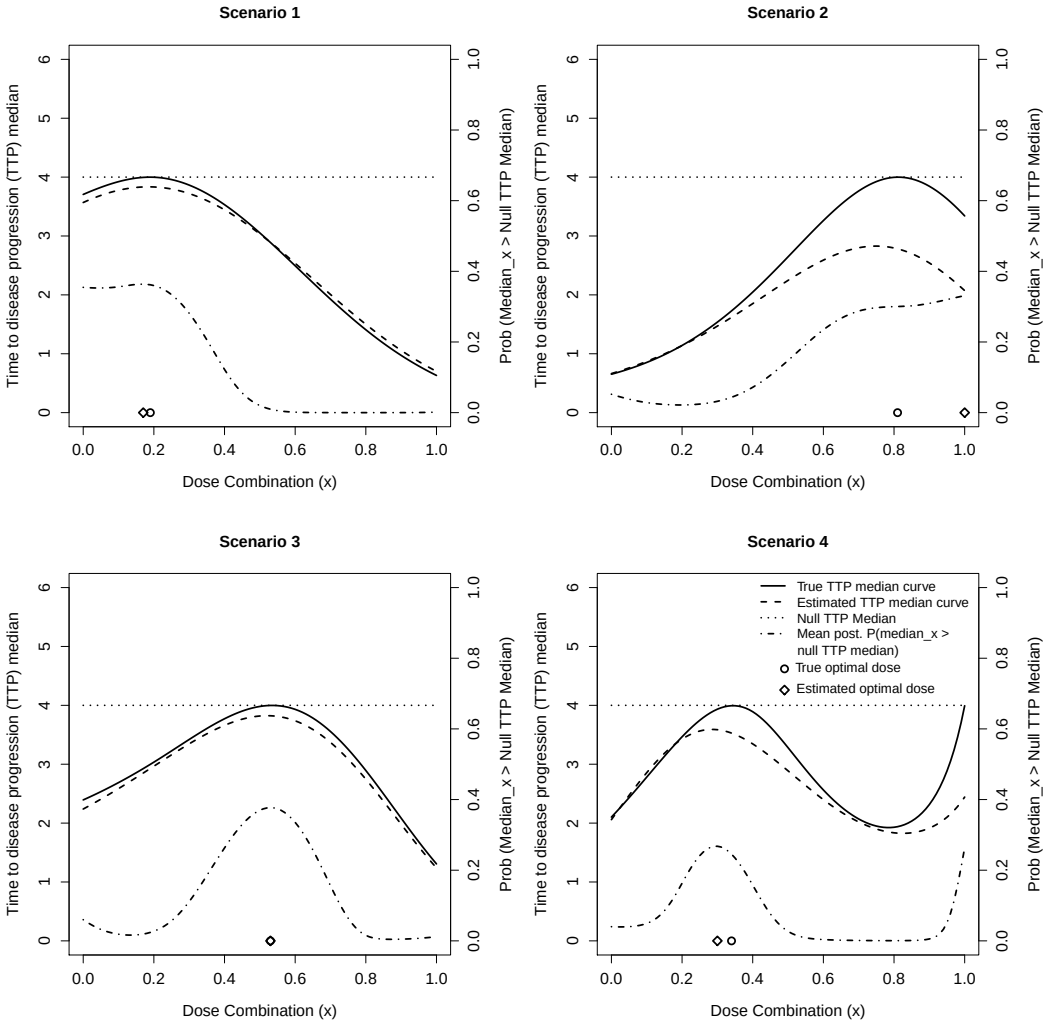


Table 3.3 Probability of early stopping under the null hypothesis in four scenarios with accrual rates of 1 and 2 patients per month.

		Probability of early stopping			Average sample size		
		δ_0			δ_0		
Scenario	Accrual rate	0.10	0.15	0.20	0.10	0.15	0.20
1	1	0.164	0.266	0.355	19.94	18.14	16.92
2		0.121	0.252	0.390	17.44	15.41	13.76
3		0.234	0.358	0.506	20.17	17.67	16.13
4		0.264	0.417	0.554	19.79	17.75	16.17
1	2	0.307	0.457	0.576	18.55	16.95	15.99
2		0.270	0.452	0.611	15.64	13.63	12.20
3		0.410	0.611	0.740	19.02	16.94	15.01
4		0.421	0.590	0.731	18.41	16.44	14.67

Appendix

In this section we display Figures that contain information regarding operating characteristics of the design and support the conclusions obtained along the manuscript.

Fig. 3.6 Pointwise average bias for scenario 1 in the first stage of the design.

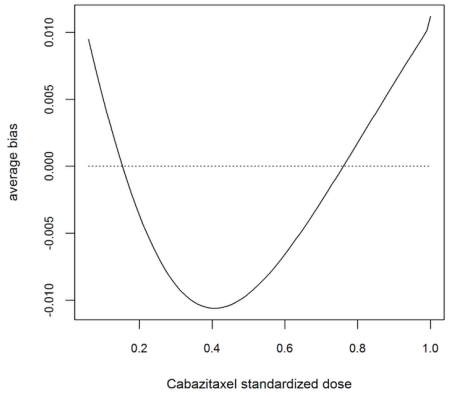


Fig. 3.7 Pointwise average bias for scenario 2 in the first stage of the design.

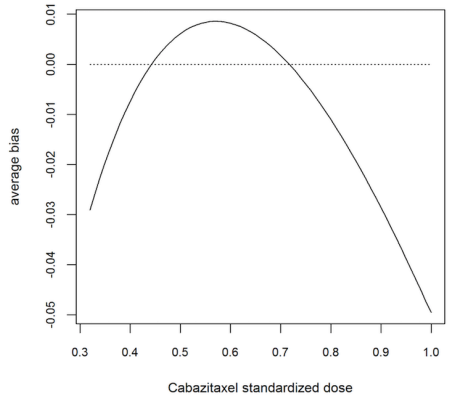


Fig. 3.8 Percent of correct recommendation for scenario 1 in the first stage of the design.

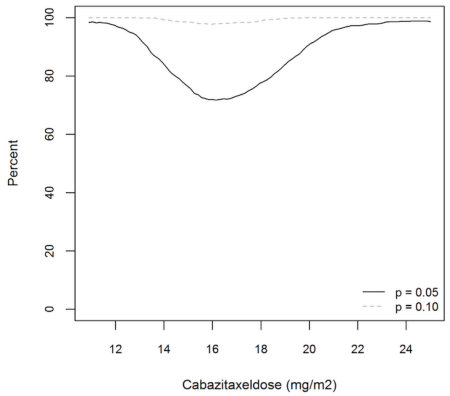


Fig. 3.9 Percent of correct recommendation for scenario 2 in the first stage of the design.

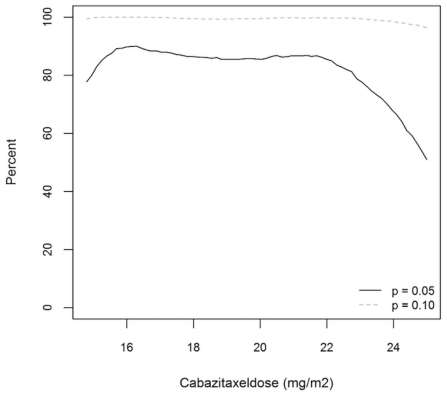


Table 3.4 Bayesian power, type I error probability and type-I + type-II error probability in two scenarios with effect size of 2 months, and accrual rates of 2 and 3 patients per month.

		Power (effect size of 2 months)		Probability of type-I error		Probability of type-I + type-II errors (effect size of 2 months)	
		δ_u		δ_u		δ_u	
Scenario	Accrual rate	0.8	0.9	0.8	0.9	0.8	0.9
1	2	0.836	0.677	0.162	0.080	0.326	0.403
2		0.742	0.577	0.121	0.052	0.379	0.475
1	3	0.824	0.639	0.167	0.081	0.343	0.442
2		0.747	0.572	0.109	0.046	0.362	0.474

Fig. 3.10 True and estimated TTP medians under four scenarios favoring the alternative hypothesis with effect size of 1.5 months and accrual rate of 2 patients per month.

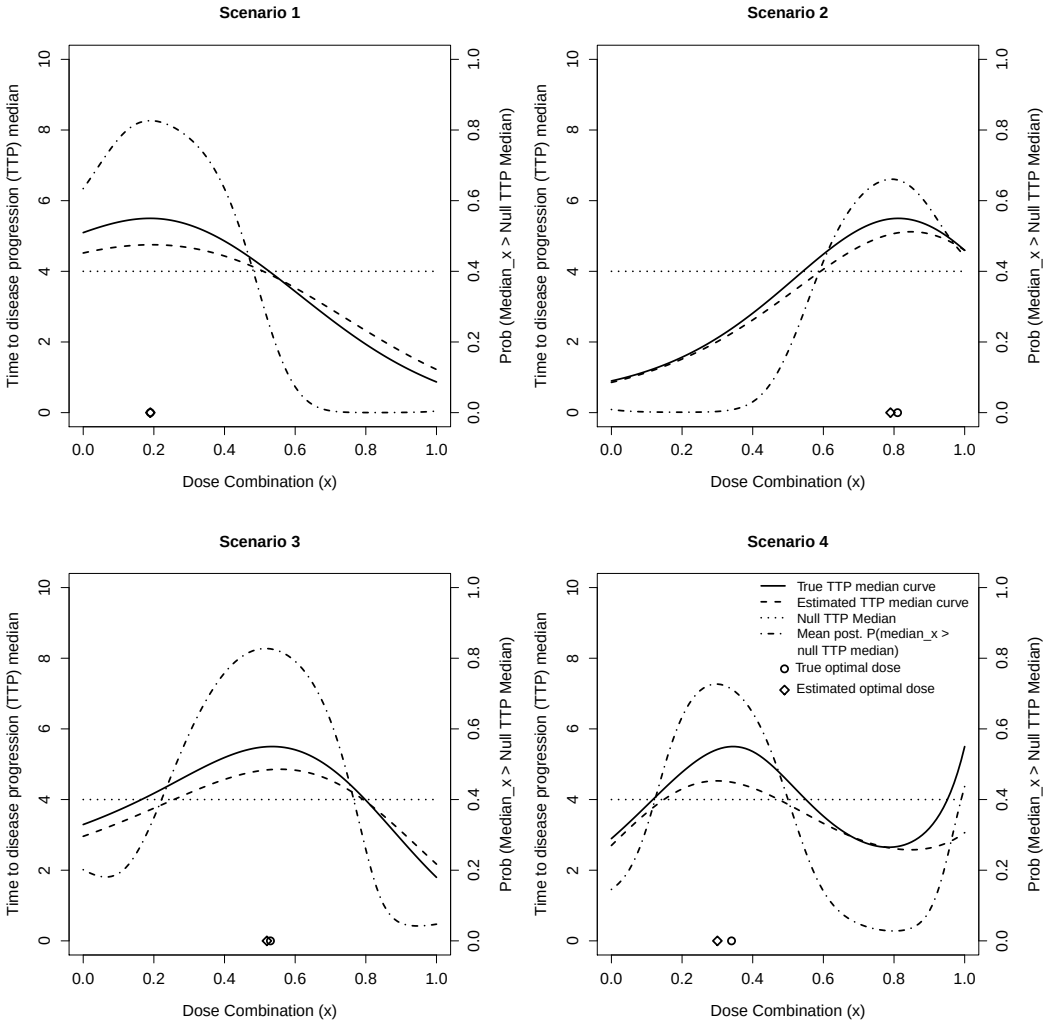


Fig. 3.11 True and estimated TTP medians under four scenarios favoring the alternative hypothesis with effect size of 2 months and accrual rate of 2 patients per month.

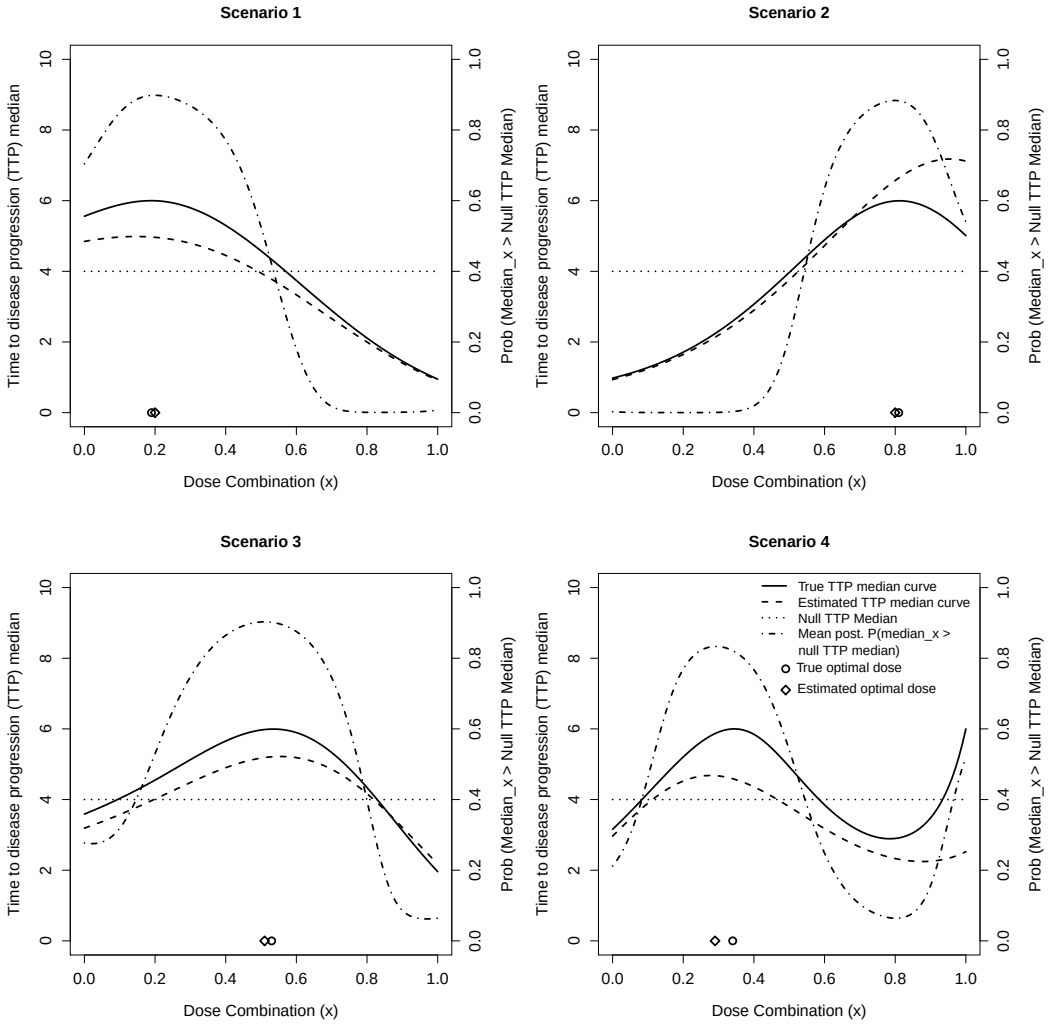


Fig. 3.12 True and estimated TTP medians under four scenarios favoring the null hypothesis with an accrual rate of 2 patient per month.

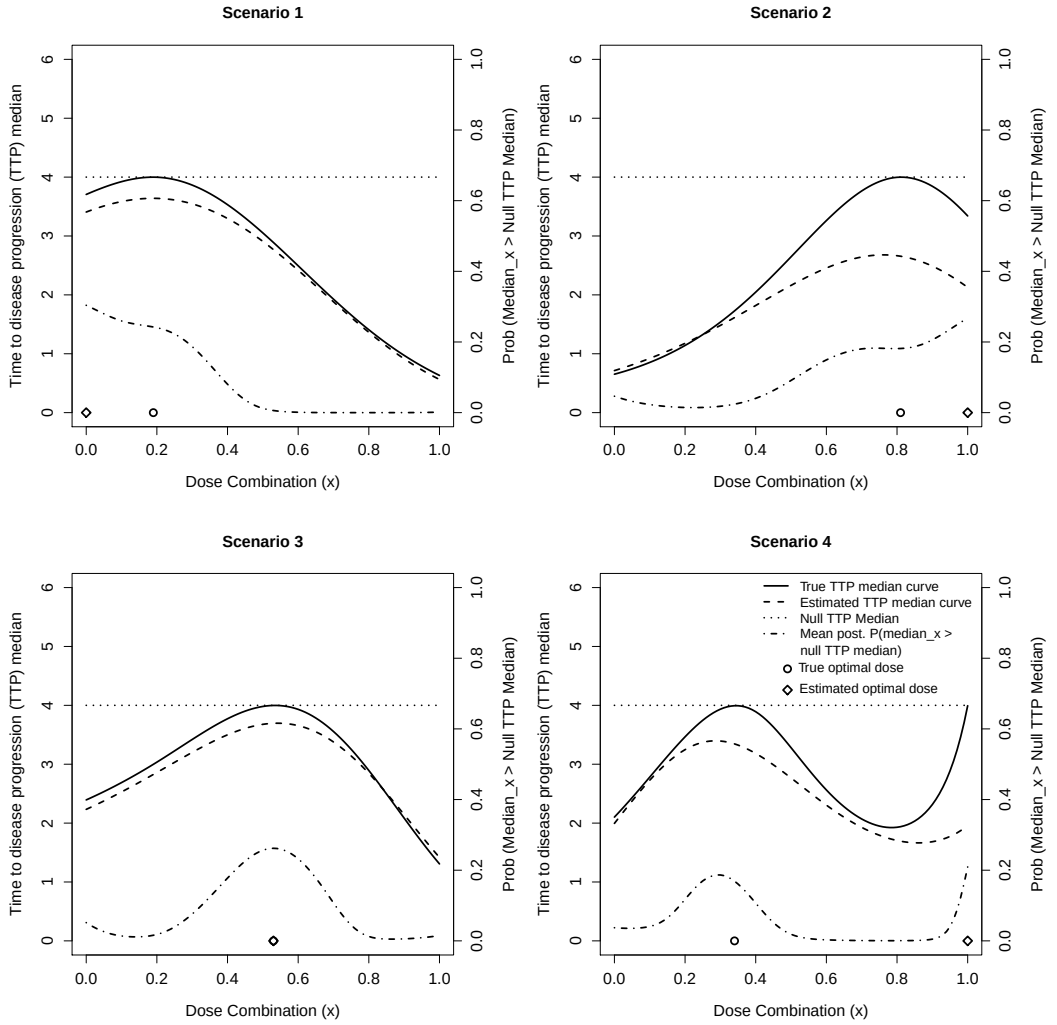
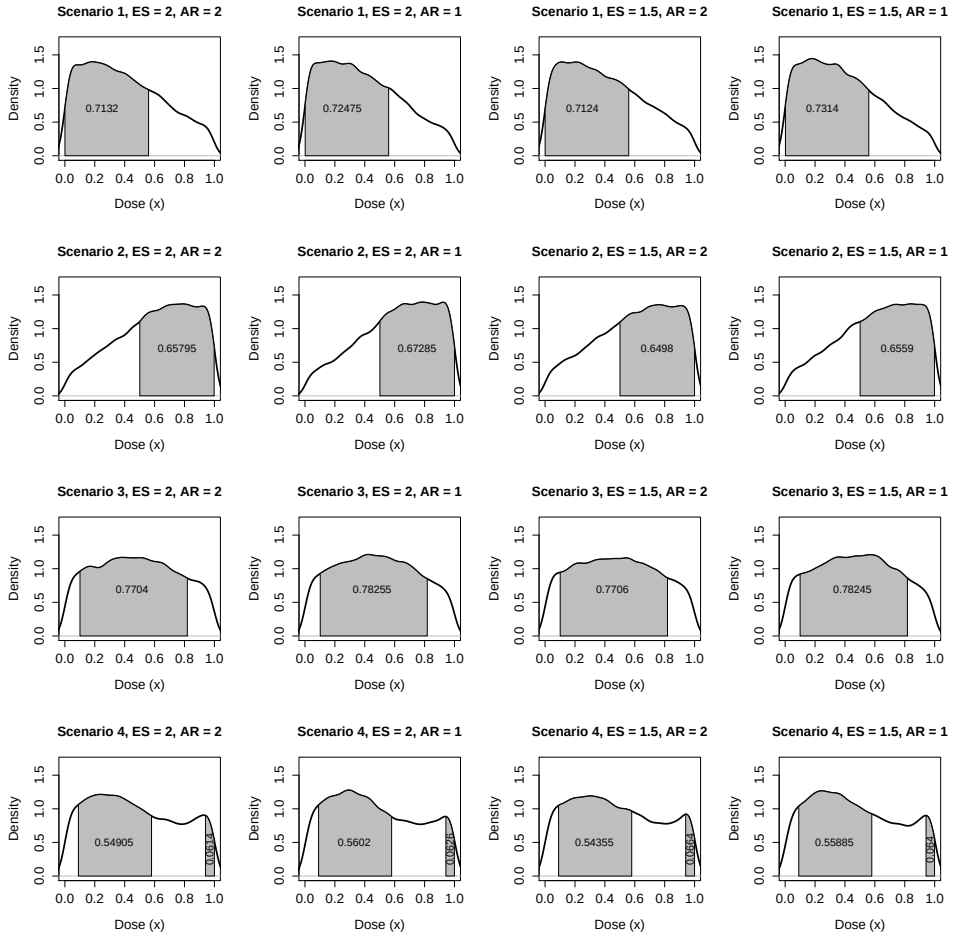


Fig. 3.13 Dose allocation density across scenarios with effect sizes (ES) of 1.5 and 2 months and accrual rates (AR) of 1 and 2 patients per months.



Chapter 4

Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects

4.1 Introduction

In drug development, randomized controlled trials remain the gold standard to confirm efficacy and safety of novel drug candidates. Often phase III trials embed formal interim analyses to allow studies to be stopped earlier for futility if the novel drug is not efficacious or for efficacy if the treatment effect is overwhelmingly positive.

Immuno-oncology (IO) is a rapidly evolving area in the development of anti-cancer drugs. IO agents can have effect on both the human immune system and the tumor microenvironment. By doing so, the tumors may be eradicated from the host or disease progression may be delayed. The effect of an IO agent is not typically directed to the tumor itself; it instead boosts or releases the brake from the patient's immune system, and this positive effect may not be observed immediately. The lag between the activation of immune cells, their proliferation and impact on the tumor is described in the literature as a delayed treatment effect. Some patients may not derive clinical benefit before their disease progresses while others may derive sustained response or control of their disease. The primary endpoints often used for confirmatory phase III studies in oncology are time to event: progression free survival (PFS) and overall survival (OS). PFS is defined as time from randomization until disease progression or death and OS is defined as time from randomization until death from any cause. The delayed treatment effect may translate to inferior or equal PFS or OS compared

to control treatment in the first months of therapy and superior survival thereafter leading to non-proportionality of hazards in the experimental and control arms of study. Therefore, the original design based on a proportional hazards assumption will lead to an underpowered study and hence both the sample size calculation and the analysis methods to be used should be reconsidered [54].

A weighted version of the log-rank test that incorporates the Fleming and Harrington class of weights [55], allows tuning the two parameters (ρ, γ) depending on if we expect early, middle or late delays, is proposed in the literature to increase the power at the end of the trial. However, tuning these parameters is not straightforward, since a misspecification may cause an even larger power drop with respect to the log-rank test.

The Fleming and Harrington class of weights, along with the estimated delay, can be incorporated into the sample size calculation in order to maintain the desired power once the treatment arm differences start to appear (see [56]).

In this article we make an empirical evaluation of the impact of having a delayed effect on power and type I error rate in the design of a confirmatory phase III study with an IO agent used in combination with a standard of care, assuming a range for delay time. We assess the performance of the weighted log-rank test as an alternative to the log-rank test given it allows weighting of late differences and the potential gain power under non-proportional hazards. The evaluation is made for both group sequential and adaptive group sequential designs with fixed values of the Fleming and Harrington class of weights. We also give some practical recommendations regarding the methodology to be used in the presence of delayed effects depending on certain characteristics of the trial.

The manuscript is organized as follows. In section 2, we describe the weighted log-rank test and derive derive the sample size calculation formula needed to incorporate the estimated delay and the Fleming and Harrington class of weights, and we introduce the combination test statistic that will be necessary when doing sample size re-assessment. In section 3 we briefly describe group sequential and adaptive group sequential designs, emphasizing two popular methods used to do sample size re-assessment. In section 4, we describe the simulated example.

4.2 Methods

In this section we describe the statistical methodology we review in this article. In sections 4.2.1 and 4.2.2 we present the weighted log-rank test and derive an optimal sample size when

using this test following [56]. This sample size derivation is presented as an alternative to the Schoenfeld's formula [57], which is normally used when calculating the necessary sample size in confirmatory trials. In section 4.2.3 we introduce the combination test statistic, which will be necessary when we perform sample size re-estimation in adaptive group sequential designs.

Let T be a vector that contains the event times, $t_i, i = 1,2,\ldots,D$, between the patients' enrollment date and the patients' final event date, t_D, such that $t_1 < t_2 < \cdots < t_D$. Let the number of events at time t_i be denoted as d_i, the total number of patients at risk at that time be denoted as n_i, and the effect delay (in months) be denoted as ε. As previously described if $t < \varepsilon$ both survival curves go in parallel and once $t \geq \varepsilon$, the survival curves will start diverging. Hence, we assume the following density functions $f_j(t)$, survival functions $S_j(t)$ and hazard functions $h_j(t)$ for the control group ($j = 1$) and for the experimental group ($j = 2$):

$$
\begin{aligned}
&f_1(t) = \lambda \exp(-\lambda t),\ S_1(t) = \exp(-\lambda t) \quad \text{and} \quad h_1(t) = \lambda, \\
&f_2(t) = \begin{cases} \lambda \exp(-\lambda t) \\ c\psi\lambda \exp(-\psi\lambda t) \end{cases},\ S_2(t) = \begin{cases} \exp(-\lambda t) \\ c\exp(-\psi\lambda t) \end{cases} \quad \text{and} \quad h_2(t) = \begin{cases} \lambda & \text{if} \quad 0 \leq t < \varepsilon \\ \psi\lambda & \text{if} \quad t \geq \varepsilon \end{cases},
\end{aligned}
\tag{4.1}
$$

where $c = \exp\left[\varepsilon\psi\lambda\left(\frac{1}{\psi-1}\right)\right]$ so that $\int_0^\infty f_2(t)\mathrm{d}t = 1$. This way, we assume a step function for the hazard ratio where from time 0 to ε, the hazard ratio is equal to 1, and from time ε the hazard ratio is equal to $1/\psi$.

In this article we assume that the control group receives the standard of care and the experimental group receives a combination of the standard of care plus the IO agent which causes the delayed effect. Hence, any observed difference from time 0 until time ε is random. The conclusions we obtain are only applicable to studies where a similar assumption is made. Otherwise, we cannot guarantee that from time 0 to time ε, both groups have a common survival function.

4.2.1 Weighted log-rank test

The weighted log-rank test is defined as

$$Z_r = \frac{\sum_{i=1}^{D} r_i(d_{1i} - \mathrm{E}(d_{1i}))}{\sqrt{\sum_{i=1}^{D} r_i^2 \mathrm{Var}(d_{1i})}}, \tag{4.2}$$

where $\mathrm{E}(d_{1i}) = n_{1i} \times \left(\frac{d_i}{n_i}\right)$, $\mathrm{Var}(d_{1i}) = \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$ and $Z_r \approx N(0,1)$ under the null hypothesis $H0 : h_1/h_2 = 1$.

[55] proposed the use of r_i to weight early, middle and late differences through the $G^{\rho,\gamma}$ class of weighted log-rank tests, where the weight function at a time point t_i is equal to

$$r_i = \hat{S}(t_i)^{\rho} (1 - \hat{S}(t_i))^{\gamma}, \tag{4.3}$$

where $\hat{S}(t_i)$ corresponds to the Kaplan-Meier estimator.

Depending on the values of ρ and γ, we will have different weight functions that will emphasize early differences $(\rho = 1, \gamma = 0)$, middle differences $(\rho = 1, \gamma = 1)$ or late differences $(\rho = 0, \gamma = 1)$ in the hazard rates or the survival curves. The parameter combination attributes equal weights to all $(\rho = 0, \gamma = 0)$ data values and hence does not emphasize any survival differences between treatment arms. Moreover, with this parameter combination (4.2) corresponds to the usual log-rank test.

As mentioned by [56], since we focus on the entire survival curve rather than the late difference, valid inference requires pre-specification of ρ and γ prior to any data collection.

Prior specification of (ρ, γ) is always advisable for the trial integrity, although some authors (see e.g., [58]) note that the value of (ρ, γ) can be modified at the interim analysis without type-I error rate inflation. At the end of the trial, we are interested in estimating the hazard ratio across the entire study, which is obtained through the standard Cox model [59]. Note however that there will be a disconnect between the hazard ratio (i.e., the standard Cox model) and the weighted log-rank test. To obtain an estimate based on the Cox model that corresponds to the weighted log-rank test see [60].

In this article we focus on the use of the weighted log-rank test in confirmatory trials with delayed effects. Other areas of use may include treatment switching, which is sometimes present in confirmatory trials and also induces non-proportional hazards (see [61]). However, it is out the scope of this article to evaluate the performance of the weighted log-rank test under the presence of treatment switching and further research on this matter would be necessary.

4.2.2 Sample size derivation for the weighted log-rank test

We introduce the optimal sample size derivation proposed by [56]. Assume that we recruit patients during time T at a certain rate in a confirmatory trial where we aim to compare survival time between two groups ($j = 1,2$): a control group, with a constant hazard over time, and an experimental group, with a hazard that changes over time. The final analysis is performed at time $T+\tau$ after the first patient is enrolled. The study period $[0, T+\tau]$ is partitioned into M subintervals of equal length $\{t_0 = 0, t_1, t_2, \ldots, t_M = T+\tau\}$. Let $h_j(t_i)$ be the hazard function for group j at time t_i and $N_j(t_i)$ be the expected number of patients at risk for group j at time t_i, where $i = 0, \ldots, M-1$.

[62] showed that the weighted log-rank statistic is normally distributed with unit variance and approximate expectation of

$$E = \frac{\sum_{i=0}^{M-1} D_i r_i \left(\frac{\phi_i \theta_i}{1+\phi_i \theta_i} - \frac{\phi_i}{1+\phi_i} \right)}{\sqrt{\sum_{i=0}^{M-1} D_i r_i^2 \frac{\phi_i}{(1+\phi_i)^2}}}, \tag{4.4}$$

where

$$\begin{aligned} &\theta_i = \frac{h_2(t_i)}{h_1(t_i)}, \;\; \phi_i = \frac{N_2(t_i)}{N_1(t_i)}, \;\; D_i = (h_1(t_i)N_1(t_i) + h_2(t_i)N_2(t_i)), \\ &N_j(t_0) = n w_j, \;\; N_j(t_{i+1}) = N_j(t_i) \left[1 - h_j(t_i) - \left(\frac{1}{T+\tau-t_i} \right) I_{\{t_i > \tau\}} \right], \end{aligned} \tag{4.5}$$

w_j represents the allocation ratio for group j, and r_i corresponds to the Fleming-Harrington's $G^{\rho,\gamma}$ class of weights where $r_i = (S(t_i))^{\rho}(1-S(t_i))^{\gamma}$ and $S(t_i)$ represents the pooled survival function. Even though it was originally proposed by [63], [56] uses $S(t_i) = w_1 S_1(t_i) + w_2 S_2(t_i)$ as a substitute for the pooled survival function, where $S_j(t_i)$ represents the survival function of group j at time t_i. However, as stated by [56], equation (4.4) can be equivalently expressed as

$$E = n^{\frac{1}{2}} E^* = n^{\frac{1}{2}} \left[\frac{\sum_{i=0}^{M-1} D_i^* r_i \left(\frac{\phi_i \theta_i}{1+\phi_i \theta_i} - \frac{\phi_i}{1+\phi_i} \right)}{\sqrt{\sum_{i=0}^{M-1} D_i^* r_i^2 \frac{\phi_i}{(1+\phi_i)^2}}} \right], \tag{4.6}$$

where

$$
\begin{aligned}
&D_i^* = (h_1(t_i)N_1^*(t_i) + h_2(t_i)N_2^*(t_i)),\\
&N_j^*(t_0) = w_j,\ \ N_j^*(t_{i+1}) = N_j^*(t_i)\left[1 - h_j(t_i) - \left(\frac{1}{T+\tau-t_i}\right)I_{\{t_i>\tau\}}\right],
\end{aligned}
\tag{4.7}
$$

Assuming that the weighted log-rank statistic is normally distributed with mean $n^{\frac{1}{2}}E^*$ and unit variance, then for a power equal to $1-\beta$ and one-sided significance level α we have

$$
\left|n^{\frac{1}{2}}E^*\right| = z_\alpha + z_\beta, \tag{4.8}
$$

where z_α and z_β correspond to the α-th and β-th percentile of the standard normal distribution respectively. The required sample size is calculated as

$$
n = \left(\frac{z_\alpha + z_\beta}{E^*}\right)^2, \tag{4.9}
$$

and the total expected number of events is equal to $n \times \sum_{i=0}^{M-1} D_i$.

4.2.3 Test statistic

We aim to test the null hypothesis, $H_0 : \frac{h_1}{h_2} = 1$, against the alternative, $H_1 : \frac{h_1}{h_2} < 1$. In the context of group sequential designs, since we are only interested in early efficacy testing we make use of the well known classical group sequential design methodology (see [64]) and make use of the O'Brien and Fleming rejection boundaries. In the context of adaptive group sequential designs, we make use of the independent increment property of the inverse normal method, which is an efficient way of incorporating data of patients who where censored at interim analysis while ensuring type-I error rate control (see [65]). The test statistic is defined as

$$
Z^* = \xi_1\Phi^{-1}(1-p_1) + \xi_2\Phi^{-1}(1-p_2), \tag{4.10}
$$

where p_1 and p_2 denote the separate stage p-values from stages 1 and 2, Φ^{-1} denotes the inverse of the standard normal distribution, and ξ_1 and ξ_2 are pre-specified weights such that $\xi_1^2 = \frac{n_1}{n_1+n_2}$, $\xi_2^2 = \frac{n_2}{n_1+n_2}$ and where n_1 and n_2 represent the number of events observed in each stage. The null hypothesis will be rejected at level α if $Z^* > \Phi^{-1}(1-\alpha)$.

However, the inverse normal method is in general not valid when doing sample size re-assessment if the adaptations depend on endpoints such OS or PFS (see [66]). We use the approach proposed by [67] where, in equation (4.10), the first stage p-value is defined by the cohort of patients included before the interim analysis and is calculated only at the end of the trial. This allows the inclusion of all the events, but it prohibits early stopping for efficacy. See [68] for a detailed review of the existing methods on this matter.

4.3 Group sequential and adaptive group sequential designs

In this section we aim to briefly describe how group sequential and adaptive group sequential designs work. For a detailed definition and explanation of this methodology see [64].

4.3.1 Group sequential designs

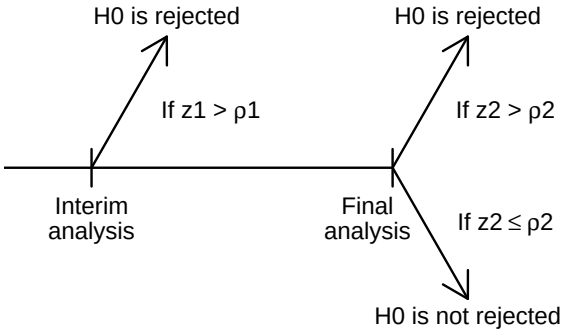
The formulae presented in section 4.2.2 allow to obtain a sample size that maintains an acceptable power at the end of the trial under the presence of delayed effects. However, a key condition is to have some knowledge about the delay of the drug. Assuming we have this knowledge when designing the confirmatory trial, we can implement a group sequential design with an interim analysis for efficacy. Note that interim analysis for futility is not advised in the presence of delayed effect because of high risk of stopping the study for futility even in scenarios that favor the alternative hypothesis.

A group sequential design with one interim analysis for efficacy is graphically described in Figure 4.1.

4.3.2 Adaptive group sequential design

Even though the sample size derivation described in section 4.2.2 guarantees that after a pre-specified effect delay we will have an acceptable power at the end of the trial while controlling the type-I error rate, we may have misspecified the delay value or maybe this value is unknown. Either way, an adaptive group sequential design that allows interim analyses and sample size re-assessment would be useful in case we expect a lack of statistical power at the end of the trial given the results at the interim analyses. Hence, with this design we aim to recover the power lost due to misspecification of the delay. As explained in section

Fig. 4.1 Graphical representation of a group sequential design with an interim analysis for efficacy where ρ_1 is the efficacy boundary at the interim analysis and ρ_2 is the efficacy boundary at the final analysis.



4.2.3, to maintain type-I error rate control when the sample size criteria is based on survival endpoints, the interim analysis is only used to do a sample size re-assessment and not for early stopping. Because we need to distinguish between the effect at the interim analysis and the effect at the final analysis, let δ_1 be the hazard ratio at the interim analysis and let δ be the hazard ratio at the end of the trial.

We now introduce two popular approaches for sample size re-assessment:

Mehta and Pocock's "promising zone" approach [69]

[69] propose a method that adaptively increases the sample size when interim results are considered "promising". For that, we compute the conditional power at the interim analysis using $\hat{\delta}_1$ rather than the true δ_1. The formula for the conditional power is defined as

$$\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) = 1 - \Phi\left(\frac{z_\alpha\sqrt{n_2} - z_1\sqrt{n_1}}{\sqrt{\tilde{n}_2}} - \frac{z_1\sqrt{\tilde{n}_2}}{\sqrt{n_1}}\right). \tag{4.11}$$

If the conditional power is within a certain pre-specified range that we consider promising, we may re-estimate the sample size to recover the power lost due to the effect delay. The selection of this range depends not only on the estimate of the effect delay but also on the budget of the sponsor for this particular trial. For example, if we have an estimated effect

delay between 3 and 7 months, but we only have budget to guarantee 80% of power up to 5 months, the sponsor can choose to stop the trial. Therefore, following [69], we partition the sample space of attainable $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2)$ values into three zones:

1. Favorable: We consider the interim results to be in the favorable zone if $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) \geq 1-\beta$. In this zone, the study is sufficiently powered for the observed $\hat{\delta}_1$ and therefore no sample size re-estimation is required.

2. Promising: We consider the interim results to be in the promising zone if $1-\beta > \mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) \geq \mathrm{CP}_{\min}$. In this zone, $\hat{\delta}_1$ is close to δ_1 but the study is not sufficiently powered and a sample size re-estimation is required. Specifically, the sample size will be increased to

$$\tilde{n}_2^*(z_1) = \min(\tilde{n}'_2, n(z_1)_{\max}), \tag{4.12}$$

   where $n_{\max}$ is the maximum sample size the sponsor is willing to enroll and $\tilde{n}'_2(z_1)$ satisfies that $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}'_2) = 1-\beta$. Following [70], it is possible to show that

$$\tilde{n}'_2 = \left(\frac{n_1}{z_1^2}\right)\left(\frac{z_\alpha\sqrt{n_2} - z_1\sqrt{n_1}}{\sqrt{n_2 - n_1}} + z_\beta\right)^2. \tag{4.13}$$

3. Unfavorable: We consider the interim results to be in the unfavorable zone if the value of $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2) < \mathrm{CP}_{\min}$. The value of $\mathrm{CP}_{\min}$ is pre-specified before the trial starts and it depends on the prior knowledge about the effect delay. In this zone the interim results are not promising and the sample size will not be re-estimated.

Type-I error rate is controlled following [71], where it is shown that the overall type-I error does not increases if the sample size is only re-assessed when

$$\mathrm{CP}_{\hat{\delta}_1}(z_1) \geq 0.5. \tag{4.14}$$

Jennison and Turnbull's "start small then ask for more" approach [72]

[72] made a detailed analysis of Mehta and Pocock's "promising zone" approach.

One drawback of the "promising zone" approach is the use of $\hat{\delta}_1$ in the construction of the promising zone and sample size increase function. The reason is that $\hat{\delta}_1$ is considered as a highly variable estimate of δ_1, and also because it is used twice in determining the

conditional power that underlies the sample size function: the first time through the value of z_1 and the second time when evaluating the conditional power at $\delta = \hat{\delta}_1$. This double use of $\hat{\delta}_1$ was also pointed out by [73] who recommends a careful inspection of the operating characteristics when using $\delta = \hat{\delta}_1$.

Another drawback of Mehta and Pocock's "promising zone" approach is that, despite the type-I error rate being controlled, because of the restriction showed in (4.14), the gain in power is relatively small for the increases in the expected sample size. Moreover, [72] demonstrated that other alternatives such us a fixed sample design and a group sequential design have exactly the same power curve and a lower expected sample size around the true value of δ.

To overcome the last limitation, [72] propose an optimal sample size calculation rule where we need to find the value of n_2^* that maximizes the objective function

$$f(n_2^*) = \text{CP}_{\hat{\delta}_1}(z_1, n_2^*) - \eta(n_2^* - n_2), \tag{4.15}$$

where η can be considered as "a tuning parameter that controls the degree to which the sample size may be increased when interim data are promising but not overwhelming".

[72] pointed out that even though the objective function given by equation (4.15) "concerns conditional probabilities given the interim data, choosing a sample size rule to optimize this objective function also yields a design with an overall optimality property expressed in terms of unconditional power". They show that

$$P_{\hat{\delta}_1}(\text{Reject } H_0) - \eta E_{\hat{\delta}}(N) = \int \{\text{CP}_{\hat{\delta}}(z_1, n_2^*(z_1)) - \eta(n_2^*(z_1) - n_2)\} f_{\hat{\delta}}(z_1) dz_1, \tag{4.16}$$

where $f_{\hat{\delta}}(z_1)$ represents the density of Z_1 under $\delta = \hat{\delta}$, and since we maximize equation (4.15), for every z_1, we also maximize the right hand side of equation (4.16). Moreover, it is possible to show that this sample size rule has the minimum expected sample size among all rules that achieve the sample power under $\delta = \hat{\delta}$.

In algorithms 1 and 2 we describe how to implement the reviewed methodology in case the sample size needs to be re-assessed.

Algorithm 1 Group sequential adaptive design using Mehta and Pocock's "promising zone" approach.

procedure
    Recruit up to n patients and when n_1 events are observed analysis compute $\mathrm{CP}_{\hat{\delta}_1}(z_1, \tilde{n}_2)$
    Calculate the number of events $\tilde{n}'_2$ and total sample size necessaries for the second stage.
    Recruit patients until $\tilde{n}'_2$ events are observed.
    Compute $Z^* = \xi_1 \Phi^{-1}(1-p_1) + \xi_2 \Phi^{-1}(1-p_2)$ ▷ p_1 is calculated at the final stage using only the patients enrolled before the interim analysis.
    if $(Z^* > Z_\alpha)$ **then**
        Outcome $\leftarrow 1$ ▷ H_0 is rejected at the final analysis
    else
        Outcome $\leftarrow 0$ ▷ H_0 is not rejected at the final analysis
    end if
    return Outcome
end procedure

Algorithm 2 Group sequential adaptive design with one interim analysis for efficacy using Jennison and Turnbull's "start small then ask for more" approach.

procedure
    Recruit up to n patients, and when n_1 events are observed do the interim analysis.
    Calculate the number of events n_2^* and total sample size necessaries for the second stage.
    Recruit patients until n_2^* events are observed.
    Compute $Z^* = \xi_1 \Phi^{-1}(1-p_1) + \xi_2 \Phi^{-1}(1-p_2)$ ▷ p_1 is calculated at the final stage using only the patients enrolled before the interim analysis.
    if $(Z^* > Z_\alpha)$ **then**
        Outcome $\leftarrow 1$ ▷ H_0 is rejected at the final analysis
    else
        Outcome $\leftarrow 0$ ▷ H_0 is not rejected at the final analysis
    end if
    return Outcome
end procedure

4.4 Simulation setup

We implement the methodology described in sections 4.2 and 4.3 on a scenario that tries to imitate a realistic phase III trial with delayed effects in oncology.

Survival data for the control arm is simulated using an exponential distribution while data for the experimental arm is simulated using a distribution that is piece-wise exponential (see equation (4.1)). Under proportional hazards, we assume that the control arm has a median survival of 6 months while the experimental arm has a median survival of 9 months. Hence, the hazard ratio is equal to 0.667. However, under the presence of delayed effects we assume a step function for the hazard ratio where it will be equal to 1 until a certain time point ε, and then it will be at its full effect after ε. This means that while the control arm will keep its median survival of 6 months, the median survival of the experimental arm will no longer be 9 months because of the delayed effect.

We establish a total study duration of 25 months, a total enrollment period of 17.5 months, randomization ratio 1:1, a power of 90% and a one-sided level α of 2.5%.

Clinical trial enrollment follows a Poisson distribution with rate of 10 patients per month. Plotting the cumulative distribution function of a Poisson distribution of these characteristics using, for instance, the R function `ecdf()`, it is straightforward to see that after 17.5 months almost all the patients, if not all, are enrolled in the trial. Results are obtained running 200,000 simulated trials. R code is showed in the appendix explaining how to simulate survival data under the presence of delayed effects.

In Table 4.1 we show the information fraction, the cumulative α spent, the O'Brien and Fleming efficacy boundaries, and the boundary crossing probability at each look. Recall that these boundaries are only used in the context of group sequential designs where the sample size is not re-assessed and they are calculated based on the information fraction only. If the sample size needs to be re-assessed, we employ different methodology (see section 4.2.3)

Table 4.1 Information fraction, the cumulative α spent, the efficacy boundaries, and the boundary crossing probability at each analysis in the group sequential design we use as an example.

Look #	Information Fraction	Cumulative α spent	Efficacy boundary Z	Boundary crossing probability (incremental)
1	0.75	0.01	2.34	0.688
2	1	0.025	2.012	0.212

For both the group sequential and the adaptive group sequential designs, we estimate the empirical power and the empirical type-I error rate at the final analysis. In the context of group sequential designs, let Z_{test} be the Z-statistic obtained at the end of the trial and Z_2 be the efficacy boundary of the final analysis presented in Table 4.1. In scenarios under the alternative hypothesis, the empirical power is defined as

$$\text{Power} = \frac{1}{M}\sum_{i=1}^{m} I[Z_{test} > Z_2], \tag{4.17}$$

whereas in scenarios under the null hypothesis, (4.17) is the empirical type-I error rate. In the context of group sequential adaptive designs, in equation (4.17), Z_{test} needs to be substituted by Z^* and Z_2 needs to be substituted by Z_α in order the implement the inverse normal method described in section 4.2.3.

4.5 Results

In this section we evaluate the repercussion of delayed effects on the power and the type-I error rate in group sequential and adaptive group sequential designs. The results presented in this section are based on the simulated scenario described in section 4.4.

Because one of the purposes of this work is to make a comparison between the log-rank test and weighted log-rank test, in Table 4.2 we show, for different delay times, the required number of events and the sample size using the parameter values $(\rho = 0, \gamma = 0)$ and $(\rho = 0, \gamma = 1)$ following the formulas presented in section 4.2.2. As we can see, under proportional hazards the parameter combination $(\rho = 0, \gamma = 0)$ is more efficient since it requires 258 events whereas the parameter combination $(\rho = 0, \gamma = 1)$ requires 369 events to maintain 90% of power. However, with 5 months delay, the parameter combination $(\rho = 0, \gamma = 1)$ becomes more efficient since it requires 741 events whereas the parameter combination $(\rho = 0, \gamma = 0)$ requires 1436 events to maintain 90% of power.

Table 4.2 Sample size calculation for different effect delay times using the parameter values $(\rho = 0, \gamma = 0)$ and $(\rho = 0, \gamma = 1)$ using the sample size formulae reviewed in Section 4.2.2.

	Delay (months)	0	1	2	3	4	5
$(\rho = 0, \gamma = 0)$	# of events	258	359	492	686	986	1436
	# of patients	330	456	621	860	1228	1777
$(\rho = 0, \gamma = 1)$	# of events	369	376	406	468	578	741
	# of patients	472	478	512	587	719	917

4.5.1 Group sequential design

In Figure 4.2 we show the empirical power and type-I error rate at the final analysis for a wide range of ρ and γ combinations with the design characteristics presented in section 4.4 assuming no delayed effect in the sample size calculation. As expected, the results show that the parameter combination $(\rho = 0, \gamma = 0)$ achieves 90% of power and 2.5% type-I error at the final analysis. However, as the delay increases, we observe that power drops faster than other combinations of ρ and γ as the effect delay increases. Other combinations like $(\rho = 0, \gamma = 1)$ have less power under proportional hazards but maintain higher power as the effect delay increases. These results are expected since low values of ρ and high values of γ weight late differences, which is the situation we recreate in this simulated trial. However, combinations that weight late differences produce a slight type-I error rate inflation as we can observe in Figure 4.2, right image.

Using the methodology described in section 4.2.2, if we incorporate an estimate of the effect delay in the sample size calculation, we are able prevent the power to drop until that specified moment. This is shown in Figure 4.3, where for each delay time we calculate the sample size necessary to achieve 90% power taking the delay into account. Moreover, when correctly specifying the effect delay, we observe that not only low values of ρ and high values of γ achieve high power. However, in terms of type-I error rate, we observe the same slight type-I error rate inflation we observed in Figure 4.2 for low values of ρ and high values of γ.

To control the type-I error rate, we propose to use a similar approach as the one used by [74] in which, although in a different context, instead of calculating the sample size for $\alpha = 2.5\%$, a lower value of α is fixed so the final type-I error rate is maintained at 2.5%.

Fig. 4.2 Empirical power and type-I error for a wide range of combinations of ρ and γ at the final analyses with different effect delay times and a unique sample size calculated assuming proportional hazards. In black, the five combinations with less cumulative power loss over time, in dark grey the power loss of the log-rank test ($\rho = 0, \gamma = 0$) over time, and in light grey the power loss of the rest of the combinations.

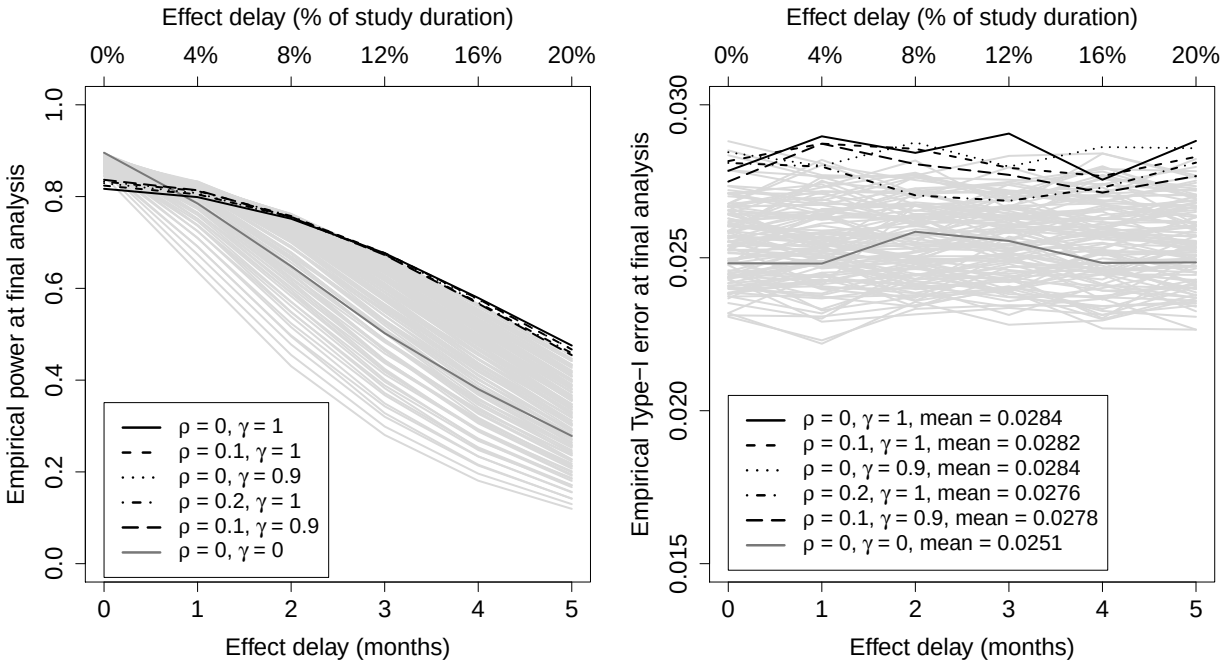
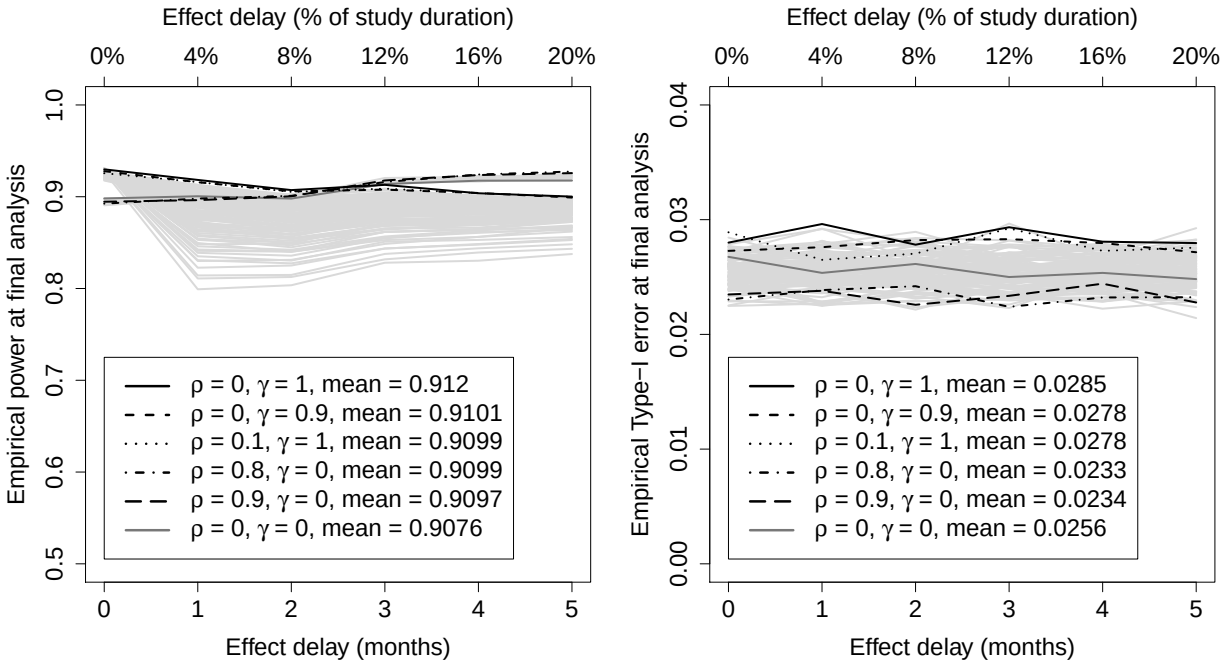


Fig. 4.3 Empirical power (left) and type-I error (right) for a wide range of combinations of ρ and γ at the final analyses with different effect delay times and a different sample size for each delay time. In the left image, in black, the five combinations with highest mean power over time. In dark grey the log-rank combination ($\rho = 0, \gamma = 0$) and in light grey the rest of the combinations. In the right image, in black the type-I error of the five combinations with highest mean power over time. In dark grey the log-rank combination ($\rho = 0, \gamma = 0$) and in light grey the rest of the combinations.



4.5.2 Adaptive group sequential design

In this section we show how performing a sample size reassessment we recover some of the power lost due to the delayed effect. As in the previous section, the results presented here make use of the simulated example described in section 4.4. However, rather than using a wide range of combinations of ρ and γ, we use the combination $(\rho = 0, \gamma = 1)$ since we believe it is the most suitable combination for this kind of setting.

In Figure 4.4 we present the empirical type-I error (top-left image), empirical power (top-right image), percent of times we re-adjust the sample size (bottom-left image) and the ratio between new sample size and original sample size (bottom-right image) for different effect delays using the weighted log-rank test with the parameter combination $(\rho = 0, \gamma = 1)$ using the promising zone approach proposed by [69].

We employ three different promising zone lower bounds (0.5, 0.1, 0.001) and compare their operating characteristics against a design that does not reassess the sample size. Without any sample size reassessment, the power is below 80% after 3 months. Using a promising zone lower bound of 0.5, the power will be below 80% after 3.5 months. However, if the promising zone lower bound is 0.1 or 0.001, the power will be below 80% after 4 and 6 months, respectively. As discussed in the literature (see [72]) we corroborate that the gains when using a lower bound of 0.5 is practically negligible and the greatest gains in power are likely to be found outside the region defined by [69].

In terms of type-I error, we observe it is perfectly controlled for any value of the promising zone lower bound. However, note that we implemented our previously described proposal in which instead of calculating the sample size for $\alpha = 2.5\%$, a lower value of α is fixed so the final type-I error rate is maintained at 2.5%. Otherwise we would see the same slight type-I error rate inflation we identified in the Figures 4.2 and 4.3 due to the ρ and γ parameters that we employ.

In terms of percent of times we fall in the promising zone, when the lower bound is 0.5, the probability of re-adjusting the sample size reaches its maximum value, which is around 15% at 4 months. If the lower bound is 0.1, the probability of re-adjusting the sample size reaches its maximum value, which is around 35% between 4 and 5 months. Last, if the lower bound is 0.001, the probability of re-adjusting the sample size reaches its maximum value, which is close to 70% at 6 months.

In terms of how much we need to increase the sample size with respect to the original sample size every time we fall in the promising zone, we observe that if the lower bound is 0.5, we need around 1.5 times the original sample size regardless the delay time. If the lower bound is 0.1, we need around 2.5 times the original sample size also regardless the delay time. Last, if the lower bound is 0.001, for a delay time $t = 0$, we need around 4.5 times the original sample size. For a delay time $t = 4$ we need around 9 times the original sample size and for a delay time $t = 6$ we need around 15 times the original sample size.

It is important to mention that, in practice, a promising zone lower bound of 0.001 may not be possible to implement given the excessively increase in the number of events needed and the consequent increase in the budget for the trial. However, we believe it is interesting to show that it is possible to maintain a power of 80% for another three extra months, regardless of the additional duration and expenses of the trial.

Last, in Figure 4.5 we make a comparison between the approaches of [69] and [72]. We selected the promising zone's lower bound 0.001 because it is the one that is more expensive to put into practice and where greater differences are observed. As expected, the approach

from [72] is able to maintain the same power as the approach from [69]. However, in terms of how much we need to increase the sample size with respect to the original sample size, [72] requires smaller sample size, specially after 4 months of delay.

Fig. 4.4 Empirical type-I error rate (top left), empirical power (top right), percent of times sample size is reassessed (bottom left) and ratio between the reassessed number of events and the original number of events (bottom right) at different delay times, when the sample size is calculated assuming no delay, using the "promising zone" approach.

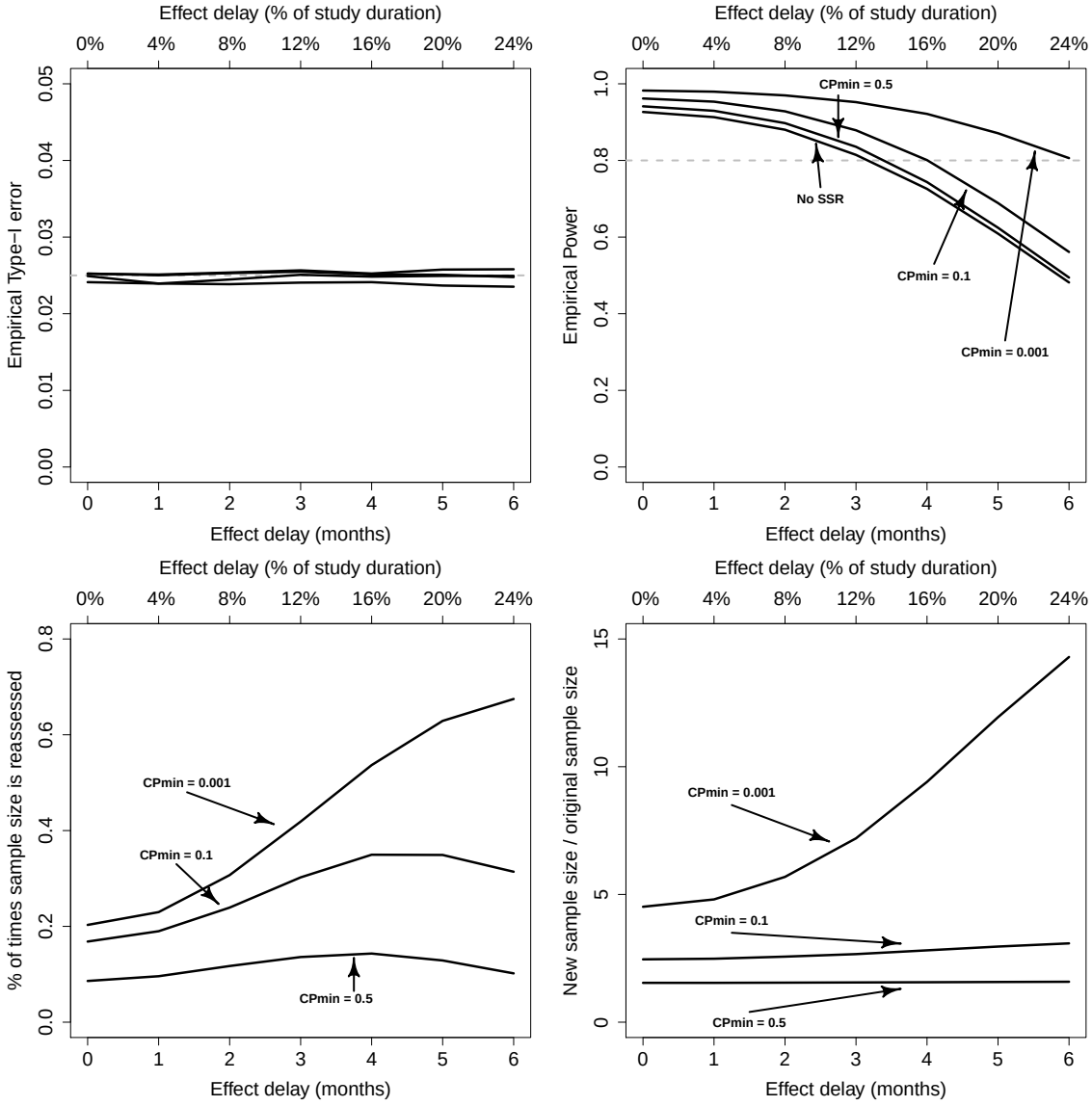
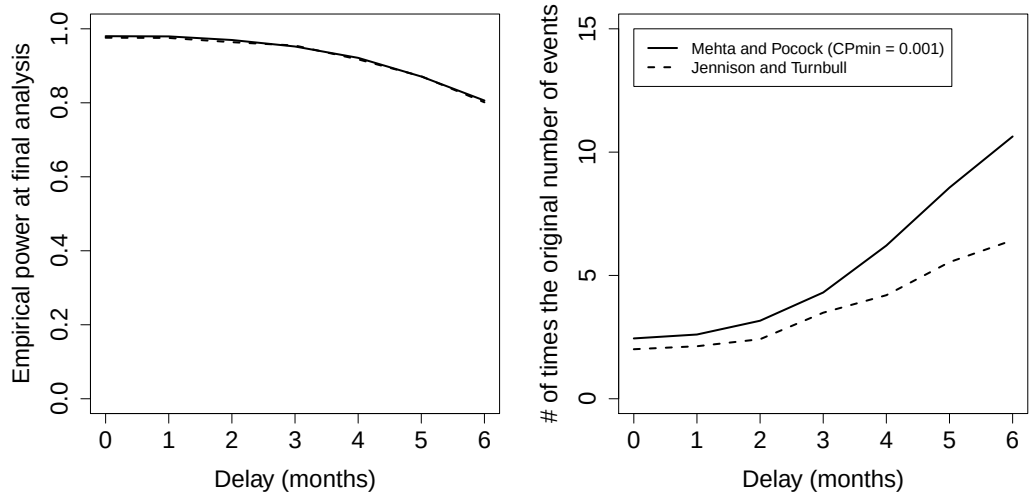


Fig. 4.5 Empirical power and ratio between the reassessed number of events and the original number of events when using the approaches from [69] and [72]



4.6 Practical Considerations

In the previous sections we evaluated the impact of delayed effects in clinical trials and what methodology exists in order to reduce it. However, we cannot conclude which methodology is better in general terms because it will depend on many factors. In this section, we emphasize some practical considerations regarding the use of the presented methodology.

The first question we tackled in this article in the use of the weighted log-rank test versus the log-rank test in group sequential and adaptive group sequential designs. In the presence of known delayed effects, we observed that the weighted log-rank test with parameter values $(\rho = 0, \gamma = 1)$ is the overall best choice, not only for the analysis but also for the sample size formula. We recall that the use of these parameter values in the weighted log-rank test generates a slight type-I error rate inflation and hence the value of α needs to be slightly decreased in order to achieve a final type-I error rate of 2.5%.

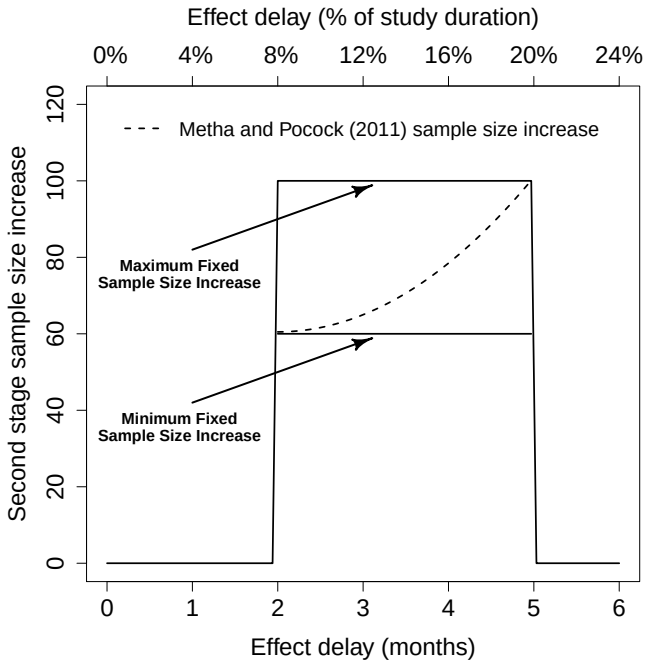
In cases where the delayed effect is unknown or underestimated in the sample size calculation, there exists methodology that re-adjusts the sample size in order to increase the power at the final analysis. The use of each method will depend on the characteristics of the trial. From the two methods we evaluated in the article, we observed that the proposal of [72] outperforms the proposal of [69] in the sense that for the same power, [72] requires less sample size. However, with these approaches it is possible to back-calculate the conditional

power at the interim analysis if we know the sample size increase recommended for the second stage of the trial. If this situation does not compromise the integrity of the trial, we recommend the use of [72] as it is proven to be more efficient. However, if the effect at the interim analysis has to remain masked, we propose the use of a modified version of [69], which would work as follows.

We would establish a promising zone, as in the original method, in which we re-calculate the sample size if the conditional power falls within a certain pre-specified range. The original method would calculate a different sample size for each conditional power (or delay time). However, in order to avoid back-calculations based on the second stage sample size, we propose to fix in advance the sample size to be used in the second stage of the trial. To avoid having an underpowered trial, we can fix the sample size increase assuming the lowest possible value for the conditional power (or the highest delay time) of the promising zone. This value would represent the maximum fixed sample size increase, although with this approach, we will be overpowering the trial for almost all values of the conditional power that fall in the promising zone. On the other hand, we can also fix the sample size increase to the highest possible value for the conditional power (or the lowest delay time) of the promising zone. This fixed value would represent the minimum fixed sample size increase, although with this approach, we will be underpowering the trial for almost all values of the conditional power that fall in the promising zone. This modification of the method proposed by [69] is illustrated (using a toy example) in Figure 4.6.

In this case, even though the "safest" option would always be using the maximum fixed sample size increase, we cannot give a recommendation since a large number of sample sizes between the maximum and the minimum fixed sample size increases can be employed and the choice depends on how much risk of having an underpowered study the sponsor is willing to take.

Fig. 4.6 Fixed sample size increase illustration following a modified version of the "promising zone" proposed by [69].



4.7 Conclusions

In this article we evaluated the impact of delayed effects, in terms of power and type-I error rate, in phase III clinical trials. We studied the use of the weighted log-rank test as an alternative to the log-rank test in group sequential and adaptive group sequential designs. This includes not only the analysis but also the incorporation of the Fleming and Harrington class of weights, as well as a delay estimate, in the sample size calculations. Also, we reviewed two different sample size re-adjustment methods, and explored which one is more efficient.

Results show that, in the presence of delayed effects when assuming proportional hazards, the weighted log-rank test with parameter values $(\rho = 0, \gamma = 1)$ was the best overall choice, as it was the one that maintained a higher power as the delay increases. When incorporating the Fleming and Harrington class of weights, as well as a delay estimate, into the sample size calculation, we observed that the power remains until the delay estimate we provided and the difference in terms of power between parameter values was not as big as under the assumption of proportional hazards, although the parameter values $(\rho = 0, \gamma = 1)$ were overall the best combination. Sample size re-adjustment allows increasing the sample size

at the interim analysis to lower the risk of failing to meet the study objective. We explored the operating characteristics of two popular approaches for sample size re-adjustment: the "promising zone" approach by [69] and the "start small then ask for more" approach by [72].

With the proposal from [69] it is possible to maintain the power high enough for the trial to be valid. However, the proposal from [72] is proven to be more efficient as for the same power curve, it requires less sample size. Nevertheless, there are situations in which having a "promising zone" may be more beneficial. This is the case when the effect at the interim analysis has to remain masked for integrity reasons. The problem is that it is possible to back-calculate the effect at the interim analysis by knowing the sample size increase. Hence, in this article we propose a modified version of the proposal from [69]. It does not require any modification of the original formulation. If a trial has a conditional power that falls in a pre-specified promising zone, we apply a pre-specified fixed sample size increase that will be used regardless the value of the conditional power as long as it falls in the promising zone. With this approach, even though we maintain the effect masked at the interim analysis, there is the risk of having an underpowered study if the fixed sample size increase in not large enough. However, if we want to avoid that risk, we will need to recruit more patients than necessary with the associated extra cost.

Chapter 5

Conclusions

5.1 Discussion

Even though this thesis focuses on adaptive designs in oncology clinical trials, it has two different parts. In the first one, we study dose finding designs in a drug combination setting. Dose finding clinical trials, also known as phases I and II, aim to identify the dose with highest efficacy among a set of doses that is considered tolerable. These two phases can be done in separate studies or in a single study with either one or two stages, depending on the indication. Phase I trials aim to find the MTD set, or in other words, a set of doses with a probability of observing DLT close to a pre-specified target, which in practice is between 0.2 and 0.4. The dose-toxicity relationship can modeled using any statistical model we think it is appropriate (e.g. logistic or copula). Dose escalation is done using mainly CRM or EWOC, which are algorithms that recommend a dose for the following cohort of patients based on previous administered doses and the target probability of DLT. As previously mentioned, phase II trials aim to identify the dose with highest efficacy with the MTD set recommended by the phase I trial. In an adaptive setting, patients are sequentially allocated to doses with high probability of efficacy based on previous administered doses. The dose-efficacy relationship can be modeled using any statistical model we think it is appropriate.

Our contributions to the field of dose finding clinical trials are:

- A Bayesian adaptive phase I trial design that allows the investigator to attribute a DLT to one or both agents in a unknown fraction of patients, even when the drugs are given concurrently [75].

- A Bayesian adaptive phase I/II design with drug combinations, a binary endpoint in stage 1, and a TTP endpoint in stage 2, where we aim to identify the dose combination region associated with the highest median TTP among doses along the MTD curve [76].

In the second part of the thesis, we focus on confirmatory clinical trials also in a drug combination setting, focusing on group sequential and adaptive group sequential designs. Confirmatory trials, or phase III trials, are the gold standard to confirm both safety and efficacy of novel drugs. These studies aim to detect differences between treatment arms while controlling the type-I error rate and maintaining a high power. One common assumption in these trials is that the hazard is not time dependent. However, in the presence of delayed effects, this assumption does not hold, and hence the design needs to be modified in order to prevent both type-I error rate inflation and power drop.

Our contribution to the field of confirmatory trials is:

- Assessment of the impact of delayed effects in group sequential and adaptive group sequential designs, and empirical evaluation in terms of power and type-I error rate of the weighted log-rank in a simulated scenario. We also give some practical recommendations regarding which methodology should be used in the presence of delayed effects depending on certain characteristics of the trial [77].

5.2 Further work

This thesis contributes to obtaining more efficient designs in phases I, II and III clinical trials. However, there is still a need of new methods in order to properly accommodate both the drug and the indication characteristics in a clinical trial. Hence, we finish the discussion by providing some ideas to extend the work presented in this thesis:

- In a phase I trial the output, that is later used in a phase II trial, is the MTD set, which is the set of doses that are considered tolerable. However, uncertainty about the MTD set is never translated into the phase II, increasing the probability of not having a successful phase II trial. Incorporating this uncertainty would increase the number of doses in the MTD set. However, given the low number of patients usually enrolled in phase II trials, there is a chance that the MTD set is too large for the sample size available and we never allocate patients in certain regions of the MTD. Adaptive

randomization methods could be used in order to make a more efficient use of the sample size available.

- Develop dose finding methodology using immuno-therapy in a drug combination setting. In a traditional phase I/II trial, unless we jointly model toxicity and efficacy, the MTD set is fixed when we move from the phase I to the phase II. Immuno-therapy is known for causing delayed effects, and hence the MTD set needs to be updated during the phase II as well because we may observe toxicity events from the phase I.

- One major question when using immuno-therapy in confirmatory trials with a group sequential design is whether at the interim analysis, there is no difference between treatment arms or the difference has not occurred yet. A possible further research line could be the use enrichment designs in order to identify and keep subgroups in which the treatment may work with higher probability.

References

[1] Ying Kuen Cheung. *Dose finding by the continual reassessment method*. CRC Press, 2011.

[2] John O'Quigley, Margaret Pepe, and Lloyd Fisher. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, pages 33–48, 1990.

[3] Sylvie Chevret. *Statistical methods for dose-finding experiments*, volume 24. John Wiley & Sons Incorporated, 2006.

[4] Scott M Berry, Bradley P Carlin, J Jack Lee, and Peter Muller. *Bayesian adaptive methods for clinical trials*. CRC press, 2010.

[5] Nolan A Wages, Mark R Conaway, and John O'Quigley. Continual reassessment method for partial ordering. *Biometrics*, 67(4):1555–1563, 2011a.

[6] Guosheng Yin and Ying Yuan. A latent contingency table approach to dose finding for combinations of two agents. *Biometrics*, 65(3):866–875, 2009a.

[7] Guosheng Yin and Ying Yuan. Bayesian dose finding in oncology for drug combinations by copula regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(2):211–224, 2009b.

[8] Nolan A Wages, Mark R Conaway, and John O'Quigley. Dose-finding design for multi-drug combinations. *Clinical Trials*, 8(4):380–389, 2011b.

[9] Yun Shi and Guosheng Yin. Escalation with overdose control for phase i drug-combination trials. *Statistics in medicine*, 32(25):4400–4412, 2013.

[10] Marie-Karelle Riviere, Ying Yuan, Frédéric Dubois, and Sarah Zohar. A bayesian dose-finding design for drug combination clinical trials based on the logistic model. *Pharmaceutical statistics*, 13(4):247–257, 2014.

[11] James Babb, André Rogatko, and Shelemyahu Zacks. Cancer phase i clinical trials: efficient dose escalation with overdose control. *Statistics in medicine*, 17(10):1103–1120, 1998.

[12] Mourad Tighiouart, Steven Piantadosi, and André Rogatko. Dose finding with drug combinations in cancer phase i clinical trials using conditional escalation with overdose control. *Statistics in medicine*, 33(22):3815–3829, 2014.

[13] Mourad Tighiouart, Quanlin Li, and André Rogatko. A bayesian adaptive design for estimating the maximum tolerated dose curve using drug combinations in cancer phase i clinical trials. *Statistics in medicine*, 36(2):280–290, 2017.

[14] Mourad Tighiouart, Yuan Liu, and André Rogatko. Escalation with overdose control using time to toxicity for cancer phase i clinical trials. *PloS one*, 9(3):e93070, 2014.

[15] Richard Simon. Optimal two-stage designs for phase ii clinical trials. *Contemporary Clinical Trials*, 10(1):1–10, 1989.

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

[17] Stuart J Pocock. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics*, pages 153–162, 1982.

[18] Peter C O'Brien and Thomas R Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979.

[19] Frank Bretz, Franz Koenig, Werner Brannath, Ekkehard Glimm, and Martin Posch. Adaptive designs for confirmatory clinical trials. *Statistics in medicine*, 28(8):1181–1217, 2009.

[20] Werner Brannath, Franz Koenig, and Peter Bauer. Multiplicity and flexibility in clinical trials. *Pharmaceutical statistics*, 6(3):205–216, 2007.

[21] Graham M Wheeler, Michael J Sweeting, Adrian P Mander, Shing M Lee, and Ying Kuen K Cheung. Modelling semi-attributable toxicity in dual-agent phase i trials with non-concurrent drug administration. *Statistics in medicine*, 36(2):225–241, 2017.

[22] Rongji Mu and Jin Xu. A new bayesian dose finding design for drug combination trials. *Statistics in Biopharmaceutical Research*, (just-accepted), 2017.

[23] Peter F Thall, Randall E Millikan, Peter Mueller, and Sang-Joon Lee. Dose-finding with two agents in phase i oncology trials. *Biometrics*, 59(3):487–496, 2003.

[24] Kai Wang and Anastasia Ivanova. Two-dimensional dose finding in discrete dose space. *Biometrics*, 61(1):217–222, 2005.

[25] Ying Yuan and Guosheng Yin. Sequential continual reassessment method for two-dimensional dose finding. *Statistics in medicine*, 27(27):5664–5678, 2008.

[26] Thomas M Braun and Shufang Wang. A hierarchical bayesian design for phase i trials of novel combinations of cancer therapeutic agents. *Biometrics*, 66(3):805–812, 2010.

[27] Adrian P Mander and Michael J Sweeting. A product of independent beta probabilities dose escalation design for dual-agent phase i trials. *Statistics in medicine*, 34(8):1261–1276, 2015.

[28] Mourad Tighiouart, Quanlin Li, Steven Piantadosi, and Andre Rogatko. A bayesian adaptive design for combination of three drugs in cancer phase i clinical trials. *American journal of biostatistics*, 6(1):1, 2016.

[29] Mauro Gasparini and Jeffrey Eisele. A curve-free method for phase i clinical trials. *Biometrics*, 56(2):609–615, 2000.

[30] John Whitehead, Helene Thygesen, and Anne Whitehead. A bayesian dose-finding procedure for phase i clinical trials based only on the assumption of monotonicity. *Statistics in medicine*, 29(17):1808–1824, 2010.

[31] Paul A Murtaugh and Lloyd D Fisher. Bivariate binary models of efficacy and toxicity in dose-ranging trials. *Communications in Statistics-Theory and Methods*, 19(6):2003–2020, 1990.

[32] Alexia Iasonos and John O'Quigley. Phase i designs that allow for uncertainty in the attribution of adverse events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(5):1015–1030, 2017.

[33] Bee Leng Lee and Shenghua Kelly Fan. A two-dimensional search algorithm for dose-finding trials of two agents. *Journal of biopharmaceutical statistics*, 22(4):802–818, 2012.

[34] David Miles, Gunter von Minckwitz, and Andrew D Seidman. Combination versus sequential single-agent therapy in metastatic breast cancer. *The Oncologist*, 7(Supplement 6):13–19, 2002.

[35] Mourad Tighiouart, Galen Cook-Wiens, and Andre Rogatko. Incorporating a patient dichotomous characteristic in cancer phase i clinical trials using escalation with overdose control. *Journal of Probability and Statistics*, 2012, 2012.

[36] Peter F Thall and Kathy E Russell. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase i/ii clinical trials. *Biometrics*, pages 251–264, 1998.

[37] Thomas M Braun. The bivariate continual reassessment method: extending the crm to phase i trials of two competing outcomes. *Controlled clinical trials*, 23(3):240–256, 2002.

[38] Anastasia Ivanova. A new dose-finding design for bivariate outcomes. *Biometrics*, 59(4):1001–1007, 2003.

[39] Peter F Thall and John D Cook. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics*, 60(3):684–693, 2004.

[40] Zhengjia Chen, Ying Yuan, Zheng Li, Michael Kutner, Taofeek Owonikoko, Walter J Curran, Fadlo Khuri, and Jeanne Kowalski. Dose escalation with over-dose and under-dose controls in phase i/ii clinical trials. *Contemporary clinical trials*, 43:133–141, 2015.

[41] Hiroyuki Sato, Akihiro Hirakawa, and Chikuma Hamada. An adaptive dose-finding method using a change-point model for molecularly targeted agents in phase i trials. *Statistics in medicine*, 35(23):4093–4109, 2016.

[42] Ying Yuan and Guosheng Yin. Bayesian dose finding by jointly modelling toxicity and efficacy as time-to-event outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(5):719–736, 2009.

[43] Xuelin Huang, Swati Biswas, Yasuhiro Oki, Jean-Pierre Issa, and Donald A Berry. A parallel phase i/ii clinical trial design for combination therapies. *Biometrics*, 63(2):429–436, 2007.

[44] Guosheng Yin, Yisheng Li, and Yuan Ji. Bayesian dose-finding in phase i/ii clinical trials using toxicity and efficacy odds ratios. *Biometrics*, 62(3):777–787, 2006.

[45] André Rogatko, Pulak Ghosh, Brani Vidakovic, and Mourad Tighiouart. Patient-specific dose adjustment in the cancer clinical trial setting. *Pharmaceutical Medicine*, 22(6):345–350, 2008.

[46] Christophe Le Tourneau, J Jack Lee, and Lillian L Siu. Dose escalation methods in phase i cancer clinical trials. *JNCI: Journal of the National Cancer Institute*, 101(10):708–720, 2009.

[47] Zhengjia Chen, Yichuan Zhao, Ye Cui, and Jeanne Kowalski. Methodology and application of adaptive and sequential approaches in contemporary clinical trials. *Journal of Probability and Statistics*, 2012, 2012.

[48] Fumiya Shimamura, Chikuma Hamada, Shigeyuki Matsui, and Akihiro Hirakawa. Two-stage approach based on zone and dose findings for two-agent combination phase i/ii trials. *Journal of biopharmaceutical statistics*, pages 1–13, 2018.

[49] Mourad Tighiouart. Two-stage design for phase i–ii cancer clinical trials using continuous dose combinations of cytotoxic agents. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2018.

[50] Mourad Tighiouart, André Rogatko, and James S Babb. Flexible bayesian methods for cancer phase i clinical trials. dose escalation with overdose control. *Statistics in medicine*, 24(14):2183–2196, 2005.

[51] Mourad Tighiouart, André Rogatko, et al. Dose finding with escalation with overdose control (ewoc) in cancer clinical trials. *Statistical Science*, 25(2):217–226, 2010.

[52] Mourad Tighiouart and Andre Rogatko. Number of patients per cohort and sample size considerations using dose escalation with overdose control. *Journal of Probability and Statistics*, 2012, 2012.

[53] A Craig Lockhart, Shankar Sundaram, John Sarantopoulos, Monica M Mita, Andrea Wang-Gillam, Jennifer L Moseley, Stephanie L Barber, Alex R Lane, Claudine Wack, Laurent Kassalow, et al. Phase i dose-escalation study of cabazitaxel administered in combination with cisplatin in patients with advanced solid tumors. *Investigational new drugs*, 32(6):1236–1245, 2014.

[54] Tai-Tsang Chen. Statistical issues and challenges in immuno-oncology. *Journal for immunotherapy of cancer*, 1(1):18, 2013.

[55] Thomas R Fleming and David P Harrington. A class of hypothesis tests for one and two sample censored survival data. *Communications in Statistics-Theory and Methods*, 10(8):763–794, 1981.

[56] Takahiro Hasegawa. Sample size determination for the weighted log-rank test with the fleming–harrington class of weights in cancer vaccine studies. *Pharmaceutical statistics*, 13(2):128–135, 2014.

[57] David A Schoenfeld. Sample-size formula for the proportional-hazards regression model. *Biometrics*, pages 499–503, 1983.

[58] John Lawrence. Strategies for changing the test statistic during a clinical trial. *Journal of biopharmaceutical statistics*, 12(2):193–205, 2002.

[59] David R Cox. Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer, 1992.

[60] Ray S Lin and Larry F León. Estimation of treatment effects in weighted log-rank tests. *Contemporary clinical trials communications*, 8:147–155, 2017.

[61] Jack Bowden, Shaun Seaman, Xin Huang, and Ian R White. Gaining power and precision by using model–based weights in the analysis of late stage cancer trials with substantial treatment switching. *Statistics in medicine*, 35(9):1423–1440, 2016.

[62] Edward Lakatos. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*, pages 229–241, 1988.

[63] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

[64] Christopher Jennison and Bruce W Turnbull. *Group sequential methods with applications to clinical trials*. CRC Press, 1999.

[65] Gernot Wassmer. Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal*, 48(4):714–729, 2006.

[66] Peter Bauer and Martin Posch. Letter to the editor. *Statistics in medicine*, 23:1333–1334, 2004.

[67] Martin Jenkins, Andrew Stone, and Christopher Jennison. An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics*, 10(4):347–356, 2011.

[68] Dominic Magirr, Thomas Jaki, Franz Koenig, and Martin Posch. Sample size reassessment and hypothesis testing in adaptive survival trials. *PloS one*, 11(2):e0146465, 2016.

[69] Cyrus R Mehta and Stuart J Pocock. Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in medicine*, 30(28):3267–3284, 2011.

[70] Ping Gao, James H Ware, and Cyrus Mehta. Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of biopharmaceutical statistics*, 18(6):1184–1196, 2008.

[71] YH Chen, David L DeMets, and KK Gordon Lan. Increasing the sample size when the unblinded interim result is promising. *Statistics in medicine*, 23(7):1023–1038, 2004.

[72] Christopher Jennison and Bruce W Turnbull. Adaptive sample size modification in clinical trials: start small then ask for more? *Statistics in medicine*, 34(29):3793–3810, 2015.

[73] Ekkehard Glimm. Comments on 'adaptive increase in sample size when interim results are promising: A practical guide with examples' by cr mehta and sj pocock. *Statistics in medicine*, 31(1):98–99, 2012.

[74] Daniel Golkowski, Tim Friede, and Meinhard Kieser. Blinded sample size re-estimation in crossover bioequivalence trials. *Pharmaceutical statistics*, 13(3):157–162, 2014.

[75] Jose L Jimenez, Mourad Tighiouart, and Mauro Gasparini. Cancer phase i trial design using drug combinations when a fraction of dose limiting toxicities is attributable to one or more agents. *Biometrical Journal*, 2017.

[76] Jose L Jimenez, Sungjin Kim, and Mourad Tighiouart. A bayesian two-stage adaptive design for cancer phase i/ii trials with drug combinations. *arXiv preprint arXiv:1809.04348*, 2018.

[77] Jose L Jimenez, Viktoriya Stalbovskaya, and Byron Jones. Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects. *arXiv preprint arXiv:1806.11294*, 2018.