5-2018

Bayesian Designs for Early Phase Clinical Trials with Novel Target Agents

Heng Zhou

BAYESIAN DESIGNS FOR EARLY PHASE CLINICAL TRIALS WITH NOVEL TARGETED AGENTS

by

Heng Zhou, M.S.

APPROVED:

Ying Yuan, Ph.D.
Advisory Professor

Yu Shen, Ph.D.

Jing Ning, Ph.D.

Yisheng Li, Ph.D.

Giulio F. Draetta, M.D., Ph.D.

APPROVED:

Dean, The University of Texas
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

BAYESIAN DESIGNS FOR EARLY PHASE CLINICAL TRIALS WITH NOVEL TARGETED AGENTS

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

by

Heng Zhou, M.S.

Houston, Texas

May, 2018

To my dear family

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest appreciation to my advisor, Dr. Ying Yuan, for his patience, guidance, and encouragement. He has been helping me not only in academia, but also in mental perspective, through this four years, from which I benefit enormously and I would take this to my future career. I would also like to thank Dr. Yu Shen and Dr. Jing Ning, for their great help and instructions in my rotations and the committee meetings. I would also like to thank other committee members, Dr. Brian Hobbs, Dr. Yisheng Li, and Dr. Giulio Draetta, for their valuable help and suggestions.

I would like to thank all my classmates and colleagues from MD Anderson and UT School of Public Health. I am blessed to be surrounded by all of them through these years, and it is a great experience that we can learn from each other and grow better and better together.

My special thanks go to my parents, for their tremendous support and understanding in my whole life.

ABSTRACT

BAYESIAN DESIGNS FOR EARLY PHASE CLINICAL TRIALS WITH NOVEL TARGETED AGENTS

Heng Zhou, M.S.

Advisory Professor: Ying Yuan, Ph.D.

My dissertation mainly focus on Bayesian designs for early phase clinical trials with novel target agents. It includes three specific topics: (1) reviewing novel phase I clinical trial designs and comparing their operating characteristics; (2) Proposing a Bayesian optimal phase II clinical trial (BOP2) design with simple and complex endpoints under a unified framework; and (3) extending the BOP2 design to incorporate the durable clinical response as a primary endpoint.

A number of novel model-based and model-assisted designs have been proposed to find the maximum tolerated dose (MTD) in phase I clinical trials, but their differences and relative pros and cons are not clear to many practitioners. We review three model-based designs, including the continual reassessment method (CRM), dose escalation with overdose control (EWOC), and Bayesian logistic regression model (BLRM), and three model-assisted designs, including the modified toxicity probability interval (mTPI), Bayesian optimal interval (BOIN), and keyboard designs. We conduct numerical studies to assess their accuracy, safety and reliability, and the practical implications of various empirical rules used in some designs, such as skipping a dose and imposing overdose control. Our results show that the CRM outperforms EWOC and BLRM with higher accuracy of identifying the MTD. For the

CRM, skipping a dose is not recommended as it substantially increases the chance of overdosing patients, while providing limited gain for identifying the MTD. EWOC and BLRM appear excessively conservative. They are safe, but have relatively poor accuracy of finding the MTD. The BOIN and keyboard designs have similar operating characteristics, outperforming the mTPI, but the BOIN is more intuitive and transparent. The BOIN yields competitive performance comparable to the CRM, but is simpler to implement and free of the issue of irrational dose assignment caused by model misspecification, thereby providing an attractive approach for designing phase I trials.

We propose a flexible Bayesian optimal phase II (BOP2) design that is capable of handling simple (e.g., binary) and complicated (e.g., ordinal, nested and co-primary) endpoints under a unified framework. We use a Dirichlet-multinomial model to accommodate different types of endpoints. At each interim, the go/no-go decision is made by evaluating a set of posterior probabilities of the events of interest, which is optimized to maximize power or minimize the number of patients under the null hypothesis. Unlike most existing Bayesian designs, the BOP2 design explicitly controls the type I error rate, thereby bridging the gap between Bayesian designs and frequentist designs. In addition, the stopping boundary of the BOP2 design can be enumerated prior to the onset of the trial. These features make the BOP2 design accessible to a wide range of users and regulatory agencies, and particularly easy to implement in practice. Simulation studies show that the BOP2 design has favorable operating characteristics with higher power and lower risk of incorrectly terminating the trial than some existing Bayesian phase II designs. The software to implement the BOP2 design is freely available at `www.trialdesign.org`.

Based on the BOP2 design, we propose a BOP2-C design which jointly models the

nested efficacy endpoints and the long-term durable clinical response (cure rate) simultaneously. We use a Dirichlet-multinomial model to account for the tumor response CR, PR, SD and PD, and assume Weibull distribution on the time to disease progression in the non-cured patients. At each interim, the go/no-go decision is made based on the posterior estimation of the CR, CR/PR and cure rates, with the optimized design parameters in the posterior probability cutoffs varying with the interim sample size. The BOP2-C design can also explicitly control the type I and type II error rates. Simulation studies show that the BOP2-C design can achieve favorable operating characteristics with high accuracy in identifying the promising treatment, and it is still robust when the true distribution of time to disease progression violates the Weibull assumption.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

As the groundbreaking achievement in novel molecular targeted agents and cancer immunotherapy in recent years, novel early phase (both phases I and II) clinical trial designs have been developed in order to deal with more complex endpoints and improve design efficiency compared to traditional designs. Phase I clinical trial designs aim to identify the maximum tolerated dose (MTD) of a new drug, which is defined as the dose with a dose-limiting toxicity (DLT) probability which is closest to the target DLT probability. The 3+3 design [1] has been dominant in phase I clinical trials for decades due to its simplicity and transparency, despite its poor ability to identify the MTD and tendency to treat patients at low doses that are potentially subtherapeutic [2]. The 3+3 design and its variations are called algorithm-based designs because they use simple, prespecified rules to guide dose escalation. The extensions of 3+3 design include the "rolling-six" design [3], the biased-coin design [4] and its variations [5, 6]. Model-based designs have been proposed that improve upon the performance of algorithm-based designs. The most well known model-based design is the continual reassessment method (CRM) [7]. The CRM assumes a parametric model for the dose-toxicity curve, and then, based on the accumulating trial data, continuously updates the estimate of the curve to guide the dose

assignment and MTD selection. Various extensions of the CRM have been proposed, including dose escalation with overdose control (EWOC) [8], Bayesian logistic regression model (BLRM) [9], time-to-event CRM [10], Bayesian model averaging CRM [11], Bayesian data-augmentation CRM [12], partial order CRM [13], and bivariate CRM [14]. Cheung provides a comprehensive review of the CRM and its related methods [15]. Compared to algorithm-based designs, model-based designs typically have superior operating characteristics. However, its use in practice has been limited probably due to its requirement of repeated model-fitting, its conceptual and computational complexity, and its non-transparent approach to decision making. Recently, a new class of designs, known as model-assisted designs [16], have been proposed to combine the simplicity of algorithm-based designs with the superior performance of model-based designs. Model-assisted designs use a model for efficient decision making like model-based designs, while their dose escalation and de-escalation rules can be tabulated before the onset of a trial as with algorithm-based designs. Examples of model assisted designs include the modified toxicity probability interval (mTPI) design [17] and its variation mTPI-2 [18], Bayesian optimal interval (BOIN) design [19, 20], keyboard design [16], BOIN combination design [21] and phase I/II design [22, 23], and keyboard combination design [24]. Mu et al. [25] proposed a generalized BOIN (gBOIN) design that handle toxicity grades, binary or continuous toxicity endpoint under a unified framework.

Phase II clinical trials are usually single-arm studies aimed at estimating the efficacy of a new treatment. They are designed to warrant efficacious treatments to be sent into large-scale randomized Phase III trials [26]. The primary endpoint for phase II clinical trial is usually patient's response to the new treatment, which is a binary outcome (response/no response). Extensive statistical methods have been

developed for Phase II clinical trial designs. A fundamental feature of phase II clinical trial designs is that they allow early termination before the maximum sample size is reached if the treatment is expected to be futile given the current observed data. Numerous designs, either frequentist or Bayesian, have been developed for phase II clinical trials. Among the frequentist designs, the most well known is Simon's optimal two-stage design [27], which minimizes the expected sample size or the maximum sample size under the null hypothesis that the treatment is not effective, while controlling the type I and II error rates at desirable levels. Other related work includes Fleming's multiple-stage test [28], Ensign's optimal three-stage design [29], and Chen's optimal three-stage design [30], among others. The major limitation of such fixed-stage methods is that they have rigorous prespecified requirements on each stage. However there is often disparity between proposed design and actual trial conduct [31, 32]. Bayesian approach could offer more flexible designs by updating information only from treated patients but not related with design setups. It also allows more flexible patient enrollment mechanism. Thall and Simon (1994) proposed a Bayesian guideline of continuously monitoring the posterior probability of whether the new treatment is promising or not, which is simple and intuitive [33]. Other Bayesian phase II trial designs developed based on the posterior probability include Heitjan's design which suggested the use of a "persuasion probability" as a consistent criterion for assessing if the new treatment is promising or not [34]. Tan and Machin [35] proposed two Bayesian two-stage designs that mimic Simon's two-stage design. Lee and Liu (2008) proposed a Bayesian phase II clinical trial design based on the predictive probability instead of the posterior probability [36].

Cancer immunotherapy is a promising treatment that stimulates the immune system to battle against cancer in the human body [37, 38, 39]. It has been shown

that cancer immunotherapy can achieve desirable clinical response in multiple types of cancer, including advanced melanoma, renal cell cancer, and non-small-cell lung cancer [40, 41, 42], etc. The mainstream cancer immunotherapeutic approaches include monoclonal antibodies (mAbs), immune checkpoint inhibitors, cancer vaccines, and other non-specific immunutherapies. A handful of clinical trials have indicated that cancer immunotherapy can achieve long-term survival and durable tumor response — rather than short-term objective tumor response — in a subset of patients. Although the proportion of patients may be small, such results could give researchers more insight to identify the unknown biomarkers that predict the durable clinical response to the treatment. Therefore, such durable clinical response should be served as a desirable endpoint in clinical trials, despite that it usually occurs in a small subset of patients, and more patients would experience low response rates and high toxicities [43].

This dissertation is organized as follows. In chapter 2, we will review several novel phase I clinical trial designs, compare their operating characteristics using the traditional 3+3 design as the benchmark. We conduct numerical studies to assess their accuracy, safety and reliability, and the practical implications of various empirical rules used in some designs, such as skipping a dose and imposing overdose control. In chapter 3, we propose a flexible Bayesian optimal phase II (BOP2) design that is capable of handling simple (e.g., binary) and complicated (e.g., nested, co-primary and eff&tox) endpoints under a unified framework. In chapter 4, we propose an extension to the BOP2 design which incorporates the durable clinical response of cancer immunotherapy for phase II clinical trials. Chapter 5 is conclusion and future work.

CHAPTER 2

Novel phase I clinical trial designs

2.1 Introduction

The development of the novel phase I clinical trial designs provides practitioners an array of tools for conducting more flexible and efficient phase I trials. We seek to compare these designs to determine their differences and relative pros and cons. In addition, some designs (e.g., CRM and BLRM) suggest optional empirical rules to regulate dose escalation, such as whether dose skipping should be allowed or an overdose control rule should be applied. Based on our experience with phase I trials at the Food and Drug Administration and MD Anderson Cancer Center, we observe that some protocols impose these empirical rules, while others do not. The practical implications of these empirical rules are not clear. To fill these knowledge gaps, we reviewed several novel phase I designs, the model-based CRM, EWOC and BLRM designs, and model-assisted mTPI, BOIN, and keyboard designs, and conducted a Monte Carlo experiment (i.e., computer simulations) to compare their operating characteristics. We note that mTPI-2 ends up with the same design as the keyboard design, thus the results for keyboard design also apply to mTPI-2. Another important issue that we examined but which is largely overlooked in the existing literature is the reliability of the design, which is defined as the likelihood of

extreme problematic trial behavior occurring under a design [19], for example, the likelihood of a design overdosing more than 50% of the patients, and the likelihood of a design failing to de-escalate the dose when 2/3 or $\geq$3/6 patients had DLTs at that dose. The incidence of such extreme behavior in a trial design may be low, but is of serious practical concern when it occurs. Our study reveals some new, intriguing design behaviors that have important practical implications. For example, two designs may have similar performance in some commonly used metrics (e.g., the average number of patients treated above the MTD), but rather different likelihood of overdosing more than 50% of the patients and failing to de-escalate the dose when 2/3 or $\geq$3/6 patients had DLTs.

Several simulation studies were carried out to compare the operating characteristics of novel phase I designs, but based on a limited number of dose-toxicity scenarios (or curves). For example, Horton, Wages and Conaway [44] compared the CRM, mTPI and BOIN designs in a simulation study with 16 dose-toxicity scenarios, and Ananthakrishnan et al. [45] considered only 3 dose-toxicity scenarios with the MTD at the same dose level. As a result, these simulation studies are prone to inadvertent selection biases of the simulation scenarios and produce the results that may not represent the general performance of the designs. In this chapter, we conduct a large scale simulation study that includes 1000 dose-toxicity scenarios. These 1000 dose-toxicity scenarios are randomly generated using a new pseudo-uniform algorithm, recently proposed by Clertant and O'Quigley [46]. Because *a priori* that algorithm does not favor any particular dose as the MTD or a particular shape of the dose-toxicity curve, it provides a neutral and objective basis for comparison.

2.2 Methods

Before reviewing the designs, we establish notation. We use $d_1 < \cdots < d_J$ to denote the J prespecified doses of the new drug that is under investigation in the trial, p_j to denote the DLT probability that corresponds to d_j, and ϕ to denote the target DLT probability for the MTD. We use n_j to denote the number of patients who have been assigned to d_j, and y_j to denote the number of DLTs observed at d_j, $j = 1, \ldots, J$. Therefore, at a particular point during the trial, the observed data are $D = \{D_j, j = 1, \ldots, J\}$, where $D_j = (n_j, y_j)$ are the "local" data observed at dose level j. For completeness and illustrating the differences between model-based and model-assisted designs, we first briefly describe the CRM, EWOC and BLRM designs, followed by the mTPI, keyboard and BOIN designs.

2.2.1 Continual reassessment method

The CRM is a model-based dose-finding approach that assumes a parametric model for the dose-toxicity curve. As information accrues during the trial, the dose-toxicity curve is re-evaluated by updating the estimates of the unknown model parameters, and the corresponding DLT probability at each investigational dose. The current estimates for the DLT probabilities are used to determine the dose allocation for the next patient or cohort of patients. One commonly used model for the CRM is the power model (also known as the empiric model) that assumes,

$$p_j(\alpha) = a_j^{\exp(\alpha)}, \quad \text{for} \quad j = 1, \ldots, J, \tag{2.1}$$

where α is the unknown parameter and $0 < a_1 < \cdots < a_J < 1$ are prior guesses for the DLT probability at each dose. The $\{a_j, j = 1, \ldots, J\}$ often are called the "skeleton" of CRM.

Under the power model in (2.1), the likelihood function for α arises as,

$$L(\alpha \mid D) = \prod_{j=1}^{J} \left\{ a_j^{\exp(\alpha)} \right\}^{y_j} \left\{ 1 - a_j^{\exp(\alpha)} \right\}^{n_j - y_j},$$

and thus, the posterior mean estimate for p_j is calculated as,

$$\hat{p}_j = \int a_j^{\exp(\alpha)} \frac{L(\alpha \mid D)\, f(\alpha)}{\int L(\alpha \mid D)\, f(\alpha)\, d\alpha}\, d\alpha,$$

where $f(\alpha)$ denotes the prior distribution for α, e.g., $N(0, 2)$. Upon updating the posterior mean estimate of the DLT probability at each dose, the next patient or cohort of patients is assigned to the "optimal" dose with an estimated DLT probability closest to the target ϕ. That is, the next patient or cohort of patients is assigned to dose level j^* such that

$$j^* = \underset{j \in (1,\ldots,J)}{\operatorname{argmin}} |\hat{p}_j - \phi|.$$

As illustrated in Figure 2.1a, the observation of DLTs tends to lift the dose-toxicity curve, leading to dose de-escalation; and the observation of no DLT tends to lower the dose-toxicity curve, leading to dose escalation. The trial continues in this manner until the prespecified sample size is exhausted. At that point, the MTD is selected as the dose with an estimated DLT probability closest to the target ϕ. In the original CRM, new patients are always assigned to the currently estimated "optimal" dose, which may lead to skipping untried doses. In practice, many trials impose a rule that forbids skipping doses and restricts dose escalation and de-escalation to one level at a time. In addition, a safety stopping rule is included such that the trial is terminated if $\Pr(p_1 > \phi \mid D) > 0.95$ (i.e., the lowest dose d_1 has more than 95% chance of being above the MTD). We imposed these practical rules for the CRM in our simulation study.

2.2.2 Dose escalation with overdose control (EWOC)

The EWOC is a modification of the CRM [8]. The EWOC employs a two-parameter logistic regression model to provide extra flexibility to model the dose-toxicity curve,

$$\text{logit}(p_j) = \beta_0 + \beta_1 d_j, \quad \beta_1 > 0, \quad j = 1, \ldots, J, \tag{2.2}$$

where β_0, β_1 are the unknown intercept and slope parameters, and d_j is the raw dosage at dose level j. To facilitate the interpretation, the EWOC reparameterizes the two-parameter logistic model using the MTD γ and the DLT probability at the first dose (i.e., p_1), as follows:

$$\gamma = \frac{1}{\beta_1}\left(\log(\phi) - \log(1-\phi) - \beta_0\right), \tag{2.3}$$

$$p_1 = \frac{\exp(\beta_0 + \beta_1 d_1)}{1 + \exp(\beta_0 + \beta_1 d_1)}. \tag{2.4}$$

The EWOC starts by treating the first cohort of patients at the lowest dose d_1. After each patient cohort is treated, the EWOC updates the estimate of the dose-toxicity curve based on the accumulating DLT data across all dose levels, and assigns the next cohort of patients to the "optimal" dose, defined as the highest dose whose posterior probability of greater than the MTD γ is equal to or less than α, i.e., $\Pr(d_j > \gamma \,|\, D) \leq \alpha$, with the recommended value of $\alpha = 0.25$. In the EWOC, dose skipping is not allowed. Thus, if the estimated optimal dose is higher than the current dose, we escalate the dose for one level; of the estimated optimal dose is lower than the current dose, we de-escalate the dose for one level. In our simulation, we used the same safety stopping rule as the CRM for the EWOC, i.e., the trial will be terminated if $\Pr(p_1 > \phi \,|\, D) > 0.95$.

2.2.3 Bayesian logistic regression model (BLRM)

The BLRM is another modification of the CRM [9]. The BLRM uses the similar two-parameter logistic regression model as the EWOC such that:

$$\text{logit}(p_j) = \log \alpha + \beta \log(\frac{d_j}{d^*}), \quad \alpha, \beta > 0, \quad j = 1, \ldots, J, \tag{2.5}$$

where α, β are unknown parameters, d_j is the raw dosage at dose level j, and d^* is the reference dosage. The BLRM requires defining the proper dosing interval (δ_1, δ_2), defined as the range of DLT probabilities regarded as acceptable. For example, given target $\phi = 0.25$, the interval (0.2, 0.3) may be defined as the proper dosing interval. The BLRM imposes an overdose control rule as follows: if the observed data suggest that there is $\geq$25 posterior probability that the DLT rate of a dose is greater than δ_2, i.e., $\Pr(p_j > \delta_2 \,|\, D) \geq 0.25$, that dose is an overdose and cannot be used to treat patients.

The BLRM starts the trial by treating the first cohort of patients at the lowest dose d_1. After each patient cohort is treated, the BLRM updates the estimate of the dose-toxicity curve based on the accumulating DLT data across all dose levels, and assigns the next cohort of patients to the optimal dose. Under the above overdose control rule, the optimal dose is defined as the dose level j that satisfies the overdose control condition $\Pr(p_j > \delta_2 \,|\, D) \geq 0.25$ and meanwhile maximizes the posterior probability of the proper dosing interval (δ_1, δ_2), i.e., $\Pr(p_j \in (\delta_1, \delta_2) \,|\, D)$. In BLRM, dose skipping is not allowed. Thus, if the estimated optimal dose is higher than the current dose, we escalate the dose for one level; of the estimated optimal dose is lower than the current dose, we de-escalate the dose for one level. The above overdose control rule leads to the following safety stopping rule: stop the trial if the lowest dose is an overdose. That is, the trial will be terminated if $\Pr(p_1 > \delta_2 \,|\, D) \geq 0.25$.

We also consider removing the overdose control rule of the BLRM, and name it BLRM-NOC. For the BLRM-NOC, the "optimal" dose is defined as the dose that maximizes the posterior probability of the proper dosing interval (δ_1, δ_2), i.e., $\Pr(p_j \in (\delta_1, \delta_2) \,|\, D)$. For the BLRM-NOC, the safety stopping rule described above cannot be used because it does not use the overdose control rule. Thus, in BLRM-NOC, we used the same safety stopping rule as the CRM, i.e., the trial will be terminated if $\Pr(p_1 > \phi \,|\, D) > 0.95$.

2.2.4 Modified toxicity probability interval (mTPI) design

The mTPI design requires the investigator to prespecify three intervals, the underdosing interval $(0, \delta_1)$, the proper dosing interval (δ_1, δ_2), and the overdosing interval $(\delta_2, 1)$. For example, given a target rate of $\phi = 0.20$, the three intervals may be defined as $(0, 0.15)$, $(0.15, 0.25)$, and $(0.25, 1)$, respectively. The mTPI design assumes

$$\begin{aligned} y_j \,|\, n_j, p_j &\sim \text{Binom}(n_j, p_j) \\ p_j &\sim \text{Beta}(1,1) \equiv \text{Unif}(0,1) \end{aligned} \tag{2.6}$$

i.e., a beta-binomial model, and thus, the posterior distribution arises as,

$$p_j \,|\, D_j \sim \text{Beta}(y_j + 1, n_j - y_j + 1), \text{ for } j = 1, \ldots, J. \tag{2.7}$$

Unlike the CRM, which models the toxicity across doses using the power model (2.1), the mTPI models toxicity only at the current dose d_j. To determine the next dose, based on D_j, the mTPI design uses the unit probability mass (UPM) corresponding

to each of the three intervals, which are defined as,

$$\begin{aligned} \text{UPM1} &= \Pr(p_j \in (0, \delta_1) \mid D_j) / \delta_1 \ , \\ \text{UPM2} &= \Pr(p_j \in (\delta_1, \delta_2) \mid D_j) / (\delta_2 - \delta_1) \ , \\ \text{UPM3} &= \Pr(p_j \in (\delta_2, 1) \mid D_j) / (1 - \delta_2) \ . \end{aligned} \tag{2.8}$$

That is, the UPM is the posterior probability that p_j lies in the corresponding interval divided by the length of that interval. Graphically, the UPM of an interval is the area under the posterior distribution curve of p_j within the interval divided by the interval length (see Figure 2.1b)

Suppose j is the current dose level. The mTPI design determines the next dose as follows:

- If UPM1 = max{UPM1, UPM2, UPM3}, then escalate the dose to level $j + 1$.
- If UPM2 = max{UPM1, UPM2, UPM3}, then stay at the current dose level j.
- If UPM3 = max{UPM1, UPM2, UPM3}, then de-escalate the dose to level $j-1$.

Because the three UPMs can be determined for all possible outcomes $D_j = (n_j, y_j)$, the dose escalation and de-escalation rules can be tabulated before the trial begins, which makes the mTPI design easy to implement in practice. The trial continues until the prespecified sample size is exhausted. At that point, the MTD is selected based on isotonic estimates of the p_j that are calculated using the pooled adjacent violators algorithm [47]. As the decision of dose escalation and de-escalation is based only on the local data at the current dose, it is possible that the dose transition oscillates between a safe dose and the next higher dose that is toxic. To avoid that issue, the mTPI design includes a dose exclusion/safety stopping rule: if $\Pr(p_j > \phi \mid n_j, y_j) > 0.95$, dose level j and higher are excluded from the trial. If the lowest dose is excluded, the trial is stopped for safety.

One drawback of using the UPM to guide dose escalation is that it lacks clear interpretation and leads to a high risk of overdosing patients [16]. To see the problem, consider a trial with a target toxicity rate of 0.20, and underdosing, proper dosing, and overdosing intervals of (0, 0.17), (0.17, 0.23), and (0.23, 1), respectively. Suppose at a certain stage of the trial, the observed data indicate that the posterior probabilities of the underdosing interval, proper dosing interval, and overdosing interval are 0.01, 0.09, and 0.9, respectively. That is, there is a 90% chance that the current dose is overdosing patients and only a 9% chance that the current dose is properly dosing patients. Despite such dominant evidence of overdosing, the mTPI design stays the same dose for treating the next patient or patient cohort, since the UPM that corresponds to the proper dosing interval is the largest. In particular, the UPM that corresponds to the proper dosing interval is $0.09/(0.23 - 0.17) = 1.5$, whereas the UPM that corresponds to the overdosing interval is $0.9/(1 - 0.23) = 1.17$.

2.2.5 Keyboard design

The keyboard design [16] resolves the overdosing issue of the mTPI by defining a series of equal-width dosing intervals (or keys) that correspond to the potential locations of the true DLT probability of a particular dose, and using the interval (or key) with the highest posterior probability to guide dose escalation and de-escalation, see Figure 2.1d. Specifically, the keyboard design starts by specifying a proper dosing interval $\mathcal{I}^* = (\delta_1, \delta_2)$, referred to as the "target key", and then populates this interval toward both sides of the target key, forming a series of keys of equal width that span the range of 0 to 1. For example, given the proper dosing interval or target key of (0.25, 0.35), on its left side, we form 2 keys of width 0.1, i.e., (0.15, 0.25) and (0.05, 0.15); and on its right side, we form 6 keys of width 0.1, i.e., (0.35, 0.45), (0.45, 0.55), (0.55, 0.65), (0.65, 0.75), (0.75, 0.85) and (0.85, 0.95). We denote the

resulting intervals/keys as $\mathcal{I}_1, \cdots, \mathcal{I}_K$. As all keys have the equal width and must be within [0, 1], some DLT probability values at the two ends (e.g., < 0.05 or > 0.95 in the example) may not be covered by keys because they are not long enough to form a key. As explained in Yan et al. [16], ignoring these "residual" DLT probabilities at the two ends does not pose any issue for decision making of dose escalation and de-escalation.

To make the decision of dose escalation and de-escalation, given the observed data $D_j = (n_j, y_j)$ at the current dose level j, the keyboard design identifies the interval $\mathcal{I}_{\max}$ that has the largest posterior probability, i.e.,

$$\mathcal{I}_{\max} = \underset{\mathcal{I}_1, \cdots, \mathcal{I}_{\mathcal{K}}}{\operatorname{argmax}} \{\Pr(p_j \in \mathcal{I}_k \mid D_j);\ k = 1, \cdots, K\},$$

which can easily be evaluated based on p_j's posterior distribution given by equation (2.7), assuming that p_j follows a beta-binomial model (2.6). $\mathcal{I}_{\max}$ represents the interval that the true value of p_j is most likely located, referred to as the "strongest" key by Yan et al. [16]. Graphically, the strongest key is the one with the largest area under the posterior distribution curve of p_j (see Figure 2.1d). If the strongest key is on the left (or right) side of the target key, that means that the observed data suggest that the current dose is most likely underdosing (or overdosing), and thus dose escalation (or de-escalation) is needed. If the strongest key is the target key, the observed data support that the current dose is most likely to be in the proper dosing interval, and thus it is desirable to retain the current dose for treating the next patient. In contrast, the UPM used by the mTPI design does not have such an intuitive interpretation and tends to distort the evidence for overdosing, as described previously.

Suppose j is the current dose level. The keyboard design determines the next dose as follows:

- If the strongest key is on the left side of the target key, then escalate the dose to level $j+1$.
- If the strongest key is the target key, then stay the current dose level j.
- If the strongest key is on the right side of the target key, then de-escalate the dose to level $j-1$.

The trial continues until the prespecified sample size is exhausted, and the MTD is selected based on isotonic estimates of p_j as described previously. During the trial conduct, the keyboard design imposes the dose exclusion/early stopping rule such that: if $\Pr(p_j > \phi \mid n_j, y_j) > 0.95$ and $n_j \geq 3$, dose level j and higher are eliminated from the trial, and the trial is terminated if the lowest dose is eliminated, where $\Pr(p_j > \phi \mid n_j, y_j)$ is evaluated based on the posterior distribution (2.7).

Similar to the mTPI design, the dose escalation and de-escalation rules of the keyboard design can be tabulated before the trial begins, making it easy to implement in practice. As the location of the strongest key approximately indicates the mode of the posterior distribution of p_j, the keyboard design can be approximately viewed as a posterior-mode-based Bayesian dose-finding method. This makes the keyboard design a new method different from the UPM-based mTPI design, despite some structural similarities between two designs (e.g., partitioning the toxicity probability into intervals and the dose escalation and de-escalation rules can be pre-tabulated). Pan, Lin and Yuan [24] showed that the keyboard design is optimal under the 0-1 loss, long-memory coherent and extended it to drug-combination trials.

2.2.6 Bayesian optimal interval (BOIN) design

Compared to the mTPI and keyboard designs, the BOIN design is more straightforward and transparent. The dose escalation and de-escalation in the BOIN

design is determined simply by comparing the observed DLT rate at the current dose with a pair of fixed dose escalation and de-escalation boundaries. Specifically, let $\hat{p}_j = y_j/n_j$ denote the observed DLT rate at the current dose, and λ_e and λ_d denote the predetermined dose escalation and de-escalation boundaries. Suppose j is the current dose level. The BOIN design determines the next dose as follows (see Figure 2.1c):

- If $\hat{p}_j \leq \lambda_e$, then escalate the dose to level $j + 1$;
- If $\hat{p}_j \geq \lambda_d$, then de-escalate the dose to level $j - 1$;
- Otherwise (i.e., $\lambda_e < \hat{p}_j < \lambda_d$), stay at the current dose level j.

The trial continues until the prespecified sample size is exhausted. At that point, select the MTD based on the isotonic estimates of DLT probabilities as described previously. During the trial conduct, the BOIN design imposes a dose elimination (or overdose control) rule as follows: if $\Pr(p_j > \phi \mid n_j, y_j) > 0.95$ and $n_j \geq 3$, dose level j and higher are eliminated from the trial, and the trial is terminated if the lowest dose is eliminated, where $\Pr(p_j > \phi \mid n_j, y_j)$ is evaluated based on the posterior distribution (2.7).

To determine the dose escalation and de-escalation boundaries (λ_e, λ_d), the BOIN design requires the investigator(s) to specify ϕ_1, which is the highest DLT probability that is deemed to be underdosing such that dose escalation is required, and ϕ_2, which is the lowest DLT probability that is deemed to be overdosing such that dose de-escalation is required. Liu and Yuan [19] provided general guidance to specify ϕ_1 and ϕ_2, and recommended default values of $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$ for general use. When needed, the values of ϕ_1 and ϕ_2 can be calibrated to achieve a particular requirement of the trial at hand. For example, if more conservative dose escalation is required, setting $\phi_2 = 1.2\phi$ may be appropriate. Given ϕ_1 and ϕ_2 and assuming

a non-informative prior (i.e., *a priori* the current dose is equally likely to be below, equal to or above the MTD), the optimal escalation and de-escalation boundaries (λ_e, λ_d) that minimizes the decision error of dose escalation and de-escalation arise as,

$$\lambda_e = \frac{\log\left(\frac{1-\phi_1}{1-\phi}\right)}{\log\left\{\frac{\phi(1-\phi_1)}{\phi_1(1-\phi)}\right\}}, \qquad \lambda_d = \frac{\log\left(\frac{1-\phi}{1-\phi_2}\right)}{\log\left\{\frac{\phi_2(1-\phi)}{\phi(1-\phi_2)}\right\}}. \tag{2.9}$$

The following table provides the dose escalation and de-escalation boundaries (λ_e, λ_d) for commonly used target DLT rate ϕ using the recommended default values $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$.

	Target DLT rate ϕ					
Boundaries	0.15	0.2	0.25	0.3	0.35	0.4
λ_e	0.118	0.157	0.197	0.236	0.276	0.316
λ_d	0.179	0.238	0.298	0.358	0.419	0.479

For example, given the target DLT rate $\phi = 0.25$, the corresponding escalation boundary $\lambda_e = 0.197$ and the de-escalation boundary $\lambda_d = 0.298$, that is, escalate the dose if the observed DLT rate at the current dose $\hat{p}_j \leq 0.197$ and de-escalate the dose if $\hat{p}_j \geq 0.298$. It has been shown that (λ_e, λ_d) are the boundaries corresponding to the Bayes factors, and thus the resulting BOIN design is optimal with desirable finite-sample and large-sample properties, i.e., long-memory coherence and consistency [19].

One interesting note is that the decision rule of the BOIN (with the non-informative prior) has an appearance of the classical frequentist design and only involves the observed DLT rate. This is common in Bayesian statistics. Many well-established Bayesian methods (e.g., estimation for normal linear regression models) result in the same estimators as the frequentist approach when non-informative pri-

ors are used. Actually, the BOIN can also be derived as a frequentist design, and its decision rule is equivalent to using the likelihood ratio test to determine dose escalation/de-escalation [19], providing another way to proof its optimality. Having both Bayesian and frequentist interpretations is a strength of the BOIN, making it appealing to wider audiences. In contrast, the mTPI and keyboard designs only have a Bayesian interpretation and require specifying priors and calculating posterior distributions.

As the observed DLT rate $\hat{p}_j$ is the most natural and intuitive estimate of p_j that is accessible by non-statisticians, the use of $\hat{p}_j$ to determine the dose escalation and de-escalation makes the BOIN design simpler and more transparent than the mTPI/mTPI-2 and keyboard designs. It is particularly easy for clinicians and regulatory agents to assess the safety of a trial using the BOIN design, thanks to the feature that the BOIN design guarantees de-escalating the dose when $\hat{p}_j \geq \lambda_d$. For example, given a target DLT rate $\phi = 0.25$, we know *a priori* that a phase I trial using the BOIN design guarantees de-escalating the dose if the observed DLT rate is higher than $\lambda_d = 0.298$ (i.e., the default value). Accordingly, the BOIN design also allows users to easily calibrate the design to satisfy a specific safety requirement mandated by regulatory agents through choosing an appropriate target DLT rate ϕ or ϕ_2. For example, supposing for a phase I trial with a new compound, the regulatory agent mandates that if the observed toxicity rate is higher than 0.25, the dose must be de-escalated. We can easily fulfill that requirement by setting the target DLT rate $\phi = 0.21$, under which the BOIN automatically guarantees de-escalating the dose if the observed toxicity rate $\hat{p}_j > \lambda_d = 0.250$. Such flexibility and transparency renders the BOIN design an important advantage in practice.

As a side note, ϕ_1 and ϕ_2 used in the BOIN design have different interpretations than the proper dosing interval (δ_1, δ_2) used in the mTPI/mTPI 2 and keyboard designs. Specifically, ϕ_1 and ϕ_2 represent the DLT rates that should be regarded as unacceptable (more precisely, underdosing and overdosing, respectively); whereas δ_1 and δ_2 represent the range of DLT probabilities that are acceptable. For example, given that the target DLT probability $\phi = 0.25$, setting $\phi_1 = 0.15$ and $\phi_2 = 0.35$ mean that the doses with the DLT rates of 0.15 and 0.35 are respectively regarded as unacceptably underdosing and overdosing, whereas setting $\delta_1 = 0.15$ and $\delta_2 = 0.35$ means that the dose with a DLT rate between 0.15 and 0.35 is regarded as acceptable. Thus, in general, the value of ϕ_1 should be smaller than δ_1 and the value of ϕ_2 should be greater than δ_2.

2.3 Software

The software for implementing the CRM is freely available at the MD Anderson Software Download Website `https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=81`. The R code for implementing the mTPI design is available at `https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=72`. The software for the BOIN design is available in three forms, including a standalone graphical user interface based Windows desktop program freely available from MD Anderson Software Download Website `https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=99`, Shiny online apps freely available at `http://www.trialdesign.org`, and R package "BOIN" available from the CRAN. The keyboard design can be implemented using the Shiny online app freely available at `http://www.trialdesign.org`.

2.4 Simulation study

2.4.1 Generating dose-toxicity scenarios

We generated true dose-toxicity scenarios using the pseudo-uniform algorithm proposed by Clertant and O'Quigley [46]. Given a target DLT probability ϕ and J dose levels, we generated scenarios as follows:

a. Select one of the J dose levels as the MTD with equal probabilities.

b. Sample $M \sim \text{Beta}(\max\{J - j, 0.5\}, 1)$, where j denotes the selected dose level, and set an upper bound $B = \phi + (1 - \phi) \times M$ for the toxicity probabilities.

c. Repeatedly sample J toxicity probabilities uniformly on $[0, B]$ until these correspond to a scenario in which dose level j is the MTD.

In these scenarios, the MTD is the dose with the DLT probability closest to the target ϕ, but not necessarily equal to the target ϕ. Consequently, it is possible to obtain scenarios in which all the doses have DLT probabilities below or above the target ϕ, as could happen in practice. If the DLT probability at the lowest dose level is greater than $\phi + 0.1$, we will claim the scenario does not have MTD, and the percentage of early termination of trials is regarded as the selection percentage of MTD. This is one of the strength of the algorithm, which provides extensive coverage on possible dose-toxicity scenarios that we may encounter in practice. Figure 2.2 displays 25 randomly selected scenarios with $\phi = 0.25$ and $J = 6$. These exhibit a variety of dose-toxicity curve shapes and spacings. The complete set of 10000 scenarios are provided in Online Appendix.

2.4.2 Simulation settings

We conducted a Monto Carlo experiment to compare the performance of the CRM, BLRM, EWOC, mTPI, BOIN, and keyboard designs, with respect to the

3+3 design. We considered three target DLT probabilities $\phi = 0.20$, 0.25 and 0.30, with 6 dose levels and a maximum sample size of 36. The starting dose level is 1. We considered 5 model-based designs: CRM (forbids dose skipping), CRM-DS (allows dose skipping), BLRM (with the overdose control rule), BLRM-NOC (with no overdose control rule), and EWOC. For the CRM, we used the `getprior(·)` function in `R` to obtain the skeleton. We set the middle dose level (i.e., dose level 3 for $J = 6$ doses) as the prior MTD, and the halfwidth of the indifference interval equal to 0.06. Specifically, when $\phi = 0.20$, the skeleton is $(0.032, 0.095, 0.20, 0.332, 0.470, 0.596)$; when $\phi = 0.25$, the skeleton is $(0.062, 0.140, 0.25, 0.376, 0.502, 0.615)$; when $\phi = 0.30$, the skeleton is $(0.095, 0.186, 0.30, 0.422, 0.540, 0.643)$. The dosages for BLRM and EWOC are (12.5, 25, 50, 100, 150, 200)mg, and the reference dosage for BLRM is $d^* = 200$mg. For BLRM, following Neuenschwander *et al.* 2008 [9], we used the vague bivariate normal distribution for the prior of $(\log \alpha, \log \beta)$, such that:

$$(\log \alpha, \log \beta) \sim N\left(\begin{pmatrix} -0.847 \\ 0.381 \end{pmatrix}, \begin{pmatrix} 2.015^2 & 0 \\ 0 & 1.027^2 \end{pmatrix}\right)$$

For EWOC, following Babb *et al.* 1998 [8], we used the non-informative priors for γ and p_1 as $\gamma \sim Unif(d_1, 2d_J - d_{J-1})$, $p_1 \sim Unif(0, \phi)$. We set the proper dosing interval $(\delta_1, \delta_2) = (\phi - 0.05, \phi + 0.05)$ for the mTPI, keyboard, BLRM and BLRM-NOC designs, and $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$ for BOIN, as recommended by these designs. The 3+3 design often completes (e.g., when 2/3 or 2/6 had DLTs) before reaching its maximum sample size. For comparability, after the 3+3 design selects the MTD, an expansion cohort is treated at the MTD to reach the total sample size of 36. Under each randomly generated scenario, we conducted 2000 simulated trials. Figure shows 25 randomly selected scenarios that display various shapes of the dose-toxicity curve. We considered cohort sizes of 3 and 1 for all designs, except

the 3+3 design. As the results are generally similar, below we focus on the cohort size of 3 with the target DLT probability of 0.25.

2.4.3 Performance metrics

For each of the 1000 scenarios, we calculated the following metrics:

A. Accuracy

A1. The percentage of correct selection (PCS), which is defined as the percentage of simulated trials in which the target dose is correctly selected as the MTD. When all the dose levels are above the MTD, PCS is defined as the percentage of early termination of trials.

A2. The average percentage of patients who are assigned to the MTD across the simulated trials. When all the dose levels are above the MTD, we use the average percentage of patients not enrolled into the trial for this metric.

B. Safety

B1. The percentage of simulated trials in which a toxic dose with the true DLT probability $\geq 33\%$ is selected as the MTD.

B2. The average percentage of patients assigned to the toxic doses with true DLT probability $\geq 33\%$.

C. Reliability

C1. The risk of overdosing, defined as the percentage of simulated trials with more than $x\%$ of patients treated at doses above the MTD. In our simulation study, we set $x\% = 50\%$, i.e., measuring the likelihood of a design assigning more than half of the patients to doses above the MTD.

C2. The risk of poor allocation, defined as the percentage of simulated trials in which fewer than 6 patients are treated at the MTD.

C3. The risk of irrational dose assignment, defined the percentage of times that the design fails to de-escalate the dose when 2/3 or $\geq$3/6 patients had DLTs at a dose.

Metrics C1 to C3 measure the likelihood of a design demonstrating extreme problematic behaviors (e.g., treating 50% or more patients at toxic doses, or fewer than 6 patients at the MTD), i.e., the reliability of the design. Although these metrics are of great practical importance, they are largely overlooked in the existing literature. Note that these reliability metrics are not covered by other metrics. For example, the percentage of patients overdosed (i.e., metric B2) does not cover the risk of overdosing (i.e., metric C1). Two designs can have a similar percentage of patients overdosed, but rather different risks of overdosing 50% of the patients (see Results). Statistically, metric B2 measures the mean of overdosing, while metric C1 measures the tail probability of overdosing. To compare the relative performance of the designs, we used the 3+3 design as a benchmark and report the difference between each of the designs and the 3+3 design for each metric. For example, the PCS for the CRM is reported as (the PCS of the CRM) $-$ (the PCS of the 3+3 design).

2.4.4 Results

A. Accuracy

Figure 2.3 A1 and A2 shows distributions of the PCS and the average percentage of patients treated at the MTD, respectively, for the investigational designs relative to the 3+3 design across 1000 scenarios. As each dose-toxicity scenario generates a value of the performance metric (e.g., PCS), we obtained a total of 1000 values for each of the metrics across the 1000 scenarios. The boxplot reflects the distribution of the metric across the 1000 scenarios. In terms of the accuracy of correctly selecting the MTD, the CRM, mTPI, BOIN and keyboard designs are comparable

and substantially outperform the 3+3 design. The BLRM and EWOC perform the worst, with the average PCS similar to that of the 3+3 design. The EWOC also has the largest variation in the PCS. The poor accuracy of the BLRM can be addressed by removing the overdose control rule: the BLRM-NOC has the highest average PCS. However, by doing so, the resulting BLRM-NOC becomes overly aggressive and treats a large percentage of patients above the MTD (as shown later). The CRM-DS, which allows dose skipping, has a slightly higher PCS than the CRM, but at the cost of increasing the risk of overdosing patients (shown later). The results for the number of patients treated at the MTD are similar to those for the PCS. The CRM, mTPI, BOIN and keyboard designs are generally comparable and substantially outperform the 3+3 design. The mTPI and CRM designs allocate slightly more patients to the MTD than the BOIN and keyboard designs, but the latter two designs are less variable, as shown by the shorter boxes in the box plot (Figure 2.3 A2). BLRM and EWOC perform the worst, and BLRM-NOC and CRM-DS perform well, with the highest average percentage of patients treated at the MTD. The EWOC is the most variable method in terms of treating patients at the MTD. To illustrate the performance of the designs under certain specific dose toxicity curves, Appendix C shows the results under 8 representative scenarios. The results are generally similar to Figure 2.3.

B. Safety

As shown in Figure 2.4 B1, the CRM, mTPI, BOIN and keyboard designs are comparable in terms of the percentage of selecting a toxic dose (with DLT probability $\geq 33\%$) as the MTD, but CRM and mTPI are slightly more variable than the BOIN and keyboard designs. BLRM-NOC not only has the highest chance of selecting a toxic dose as the MTD, but also is the most variable. The BLRM and EWOC designs

are the most conservative and least likely to select a toxic dose as the MTD. In terms of the percentage of patients treated at a toxic dose with DLT probability $\geq 33\%$, BLRM-NOC and CRM-DS stand out as the most aggressive designs, see Figure 2.4 B2. These two designs treat substantially more patients at toxic doses than the other designs and exhibit the largest variation. On average, the CRM, mTPI, BOIN and keyboard designs are comparable, but BOIN and keyboard show smaller variations.

The reason mTPI is more likely than the other designs to overdose at least 50% of the patients is explained previously (e.g., the UPM cannot appropriately measure the evidence of the toxicity of a dose), and can also be seen through the dose escalation and de-escalation rules for the three model-assisted designs reported in Table 2.1. When the target is $\phi = 0.20$, the default BOIN, mTPI and keyboard designs use different thresholds for dose escalation and de-escalation. In particular, compared to the BOIN and keyboard designs, the mTPI design is less likely to de-escalate the dose when a high rate of toxicity is observed. For example, suppose 6 patients have been treated at the current dose, the BOIN and keyboard designs will de-escalate the dose if 2 DLTs are observed, whereas the mTPI requires observing 3 DLTs before de-escalating the dose. Consequently, the mTPI design tends to stay long (i.e., get stuck) at a particular dose. If that particular dose is above the MTD, a large percentage of patients are overdosed.

C. Reliability

In terms of the risk of overdosing 50% or more of the patients (Figure 2.5 C1), the BLRM, BOIN and keyboard designs perform the best, and BLRM-NOC performs the worst, with significantly higher (i.e., about 10% higher on average) risk. The performance of the CRM and mTPI designs are similar and rank in between the performances of these other designs. The EWOC has similar averaged risk of

overdosing patients as BOIN and keyboard designs, but is much variable. We note that CRM, mTPI, BOIN and keyboard, on average, overdose similar percentages of patients (Figure 2.4 B2), but have different risks of overdosing 50% or more of the patients (Figure 2.5 C1). This indicates that the risk of overdosing (50% or more patients) and the average percentage of patients overdosed indeed measure different aspects of a design, and it is thus important to consider both metrics when evaluating a design. Compared to the CRM, CRM-DS had about 5% higher risk of overdosing 50% or more of the patients on average due to its aggressive dose skipping. In terms of the risk of poor allocation (i.e., treating fewer than 6 patients at the MTD, see Figure 2.5 C2), BLRM and EWOC perform the worst, with a significantly higher risk than the other designs. The CRM, CRM-DS, BLRM-NOC, BOIN and keyboard designs have comparable risks of poor allocation and keyboard design (thus mTPI-2 as well) improves the mTPI design.

In terms of the risk of irrational dose assignment (Figure 2.5 C3), the model-assisted designs outperform the model-based designs. The model-based designs (i.e., the CRM, BLRM and EWOC) have 8% to 55% chance of failing to de-escalate the dose when 2/3 or $\geq$3/6 patients had DLTs, whereas such irrational dose assignments never occur in mTPI, BOIN and keyboard designs. To the best of our knowledge, this result is new and no literature has studied such in-trial behavior of designs. Our result discloses a disturbing, yet unsurprising, behavior of model-based designs. The model-based designs rely on the assumed model to make the decision of dose assignment. When the model is misspecified, the estimates can be biased and thus irrational dose assignment arises. The model-assisted designs are free of that issue because they do not impose any model assumption on the dose-toxicity curve. For example, by its dose escalation/de-escalation rule, the BOIN guarantees de-escalating

the dose if the observed DLT rate at the current dose is higher than 29.8%, given the target DLT rate of 25%.

2.4.5 Analysis of simulation results

Our simulation results show that the CRM, BOIN and keyboard designs have comparable, good operating characteristics, especially in terms of PCS and the risk of overdosing a large percentage of patients. However, when we examine each scenario individually, we find that in certain scenarios the CRM has much higher PCS than the BOIN and keyboard designs, while in other scenarios the reverse is true. In this section we aim to characterize the scenarios in which the CRM outperforms the BOIN design, and vice versa. Because the keyboard design has very similar performance as the BOIN, in what follows, we focus on the CRM and BOIN. In the trial conduct of CRM, the parameter α in (2.1) is continuously updated to reflect the accruing data. We hypothesized that if there exists an α_0 such that the fitted toxicity probabilities $\boldsymbol{\pi}(\alpha_0) = (a_1^{\exp(\alpha_0)}, \ldots, a_J^{\exp(\alpha_0)})$ from the power model are close to the true toxicity probabilities (i.e., if the power model provides a good fit to the true dose-toxicity curve), then the CRM will outperform the BOIN design, and vice versa. To verify our hypothesis, first, given a specific dose-toxicity scenario with true toxicity rates $(p_1^*, \cdots, p_J^*)$, we defined a goodness-of-fit (GOF) index for the toxicity probabilities as follows,

$$GOF = \min_{\alpha} \sqrt{\sum_{j=1}^{J} \left(a_j^{\exp(\alpha)} - p_j^* \right)^2}.$$

The GOF index summaries the difference or distance between the best-fitted CRM model and the true dose-toxicity curve, in terms of the mean square error. A smaller value indicates that the CRM model can provide a better fit to the true dose-toxicity curve. The value of GOF is determined through the grid search over α. Second, we

selected the scenarios in which CRM had a PCS that was at least 10% higher than BOIN, and the scenarios in which BOIN has a PCS that was at least 10% higher than CRM. Third, using K-means clustering [48], we partitioned these two sets of scenarios into 3 clusters. Figure 2.6 shows the median scenario in each cluster, as well as the best model-fitted curve (dotted lines). Figure 2.6 shows that, compared to the dose-toxicity curves that favor the BOIN design, the dose-toxicity curves that favor the CRM are closer to the corresponding best model-fitted curve. The first three scenarios—in which the CRM has better PCS than BOIN—correspond to smaller GOF indexes than the latter three scenarios—in which BOIN has better PCS than the CRM.

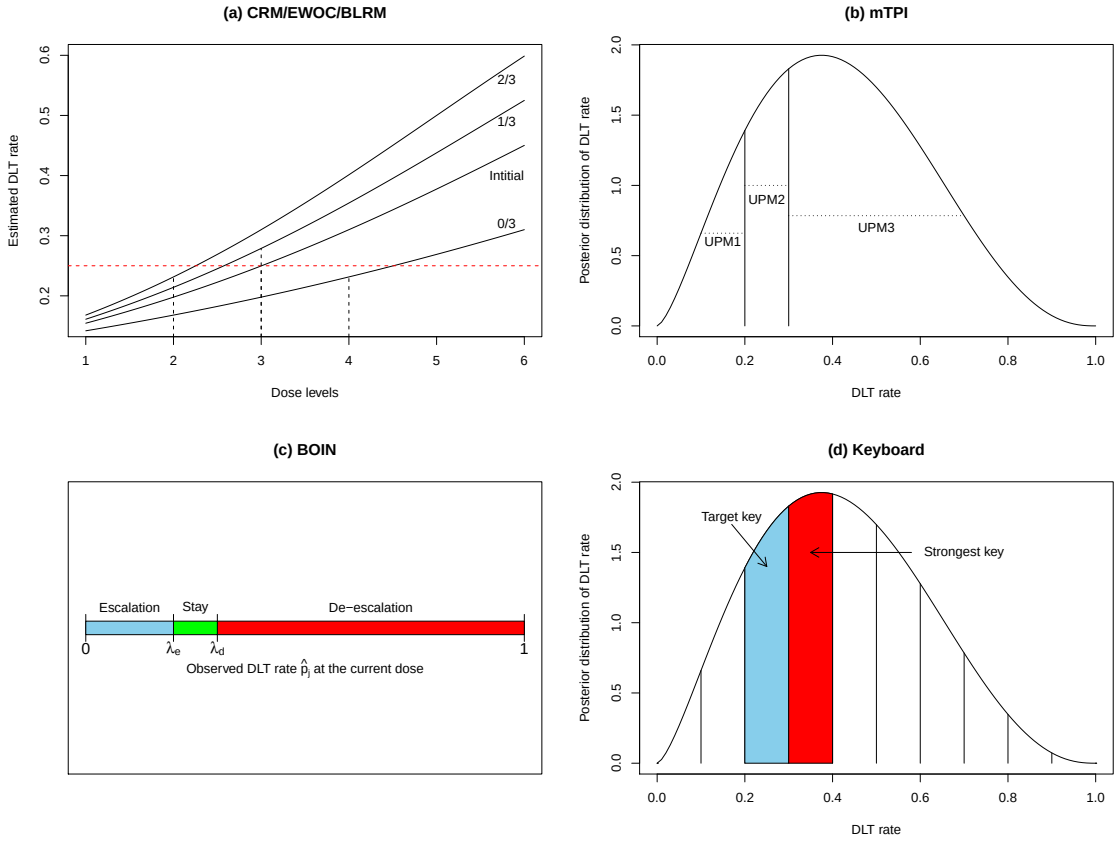

Figure 2.1: Decision of dose escalation and de-escalation under the CRM/EWOC/BLRM, mTPI, BOIN and keyboard designs. (a) CRM/EWOC/BLRM uses the estimated dose-toxicity curve that is continuously updated based on accumulative data; curve labeled initial is the initial estimate of the dose-toxicity curve before the first cohort is treated; curve labels "0/3", "1/3" and "2/3" represent the updated estimate of the dose-toxicity curve when 0/3 and 1/3 and 2/3 patients had DLT, respectively. (b) mTPI calculates and compares the UPMs of the underdosing, proper dosing and overdosing intervals. (c) BOIN compares the observed DLT rate at the current dose with the prespecified dose escalation boundary λ_e and de-escalation boundary λ_d. (d) The keyboard design forms a series of equal-width keys and bases the decision on the position of the strong key with respect to the target key.

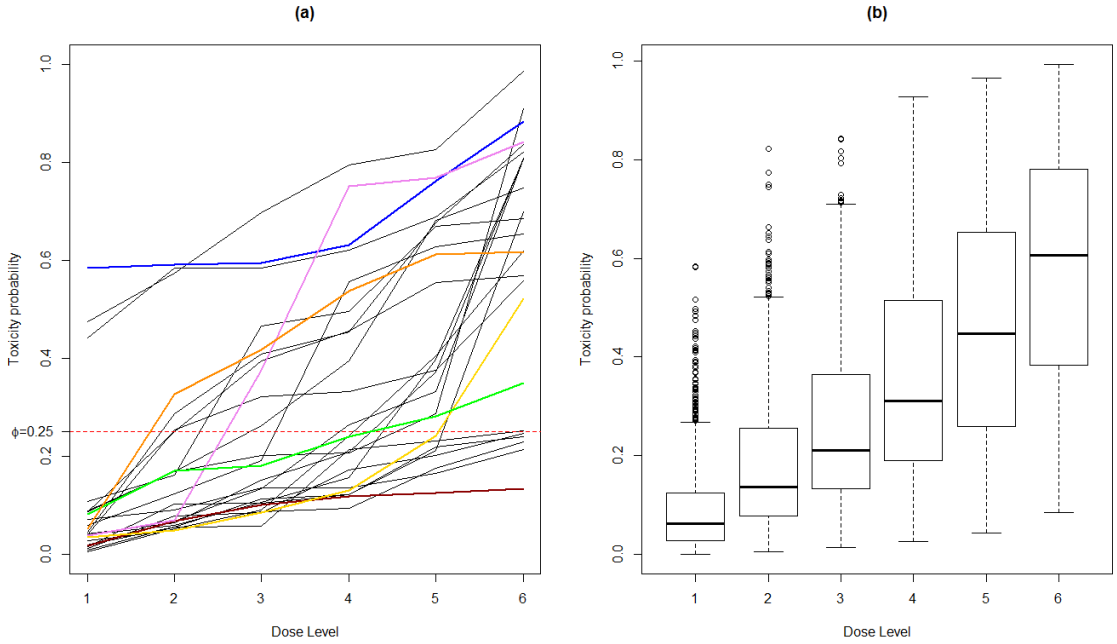

Figure 2.2: Relation between dose level and toxicity. (a) 25 randomly selected dose-toxicity curves with 6 picked curves showing different shapes; (b) distribution of the DLT probabilities by dose level from the 1000 scenarios.

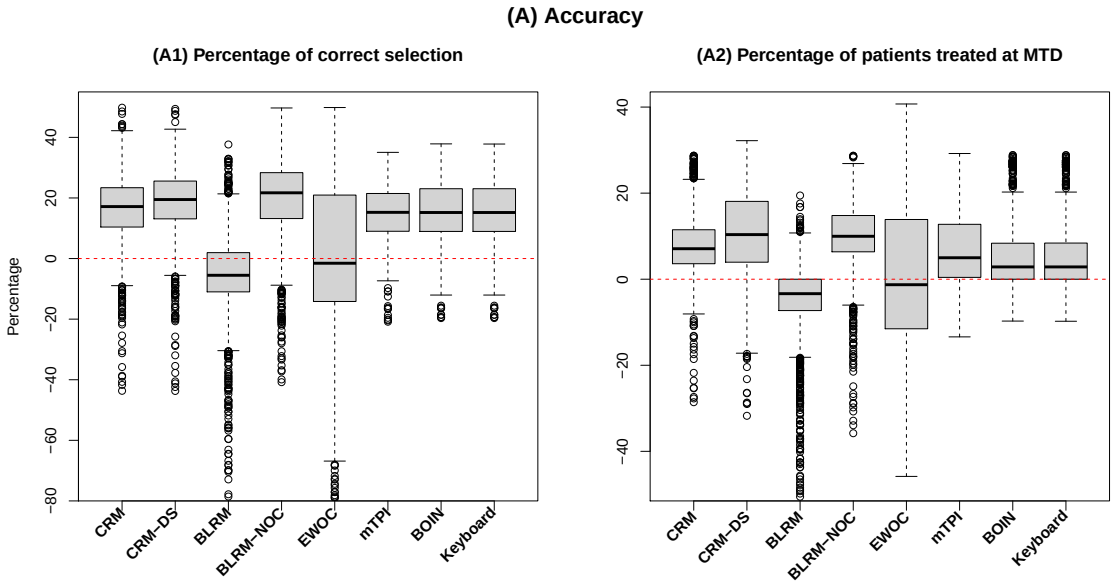

Figure 2.3: Comparison of accuracy metrics for the 8 designs with respect to the 3+3 design. A1. Percentage of correct selection of the MTD; A2. Percentage of patients treated at the MTD; A larger value indicates better performance; positive value means that the design outperforms the 3+3 design.

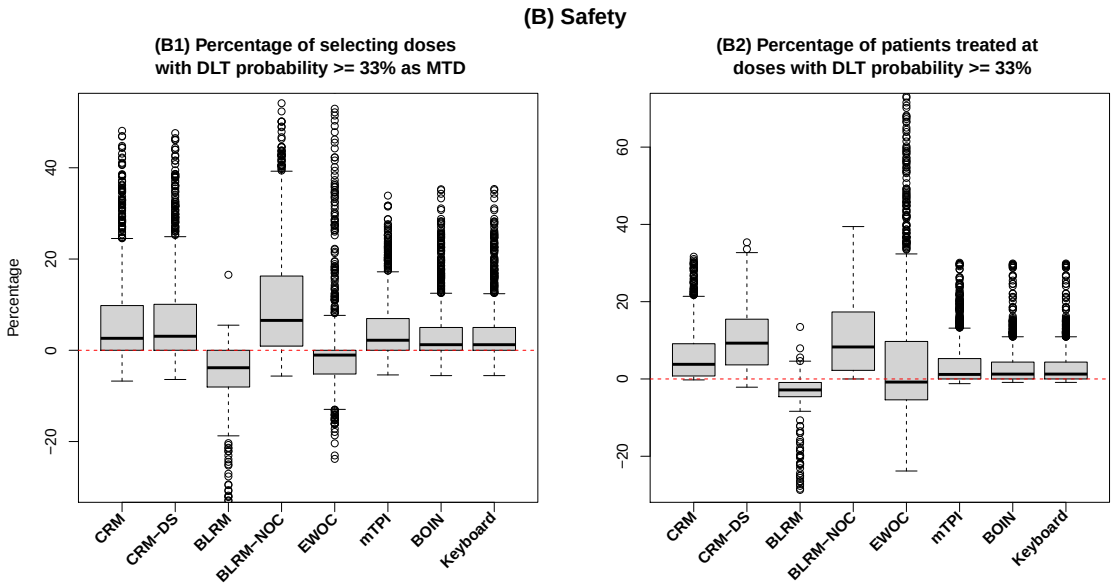

Figure 2.4: Comparison of safety metrics for the 8 designs with respect to the 3+3 design. B1. Percentage of selecting doses with DLT probability $\geq 33\%$ as the MTD; B2. Percentage of patients treated at doses with DLT probability $\geq 33\%$; A smaller value indicates better performance; negative value means that the design outperforms the 3+3 design.

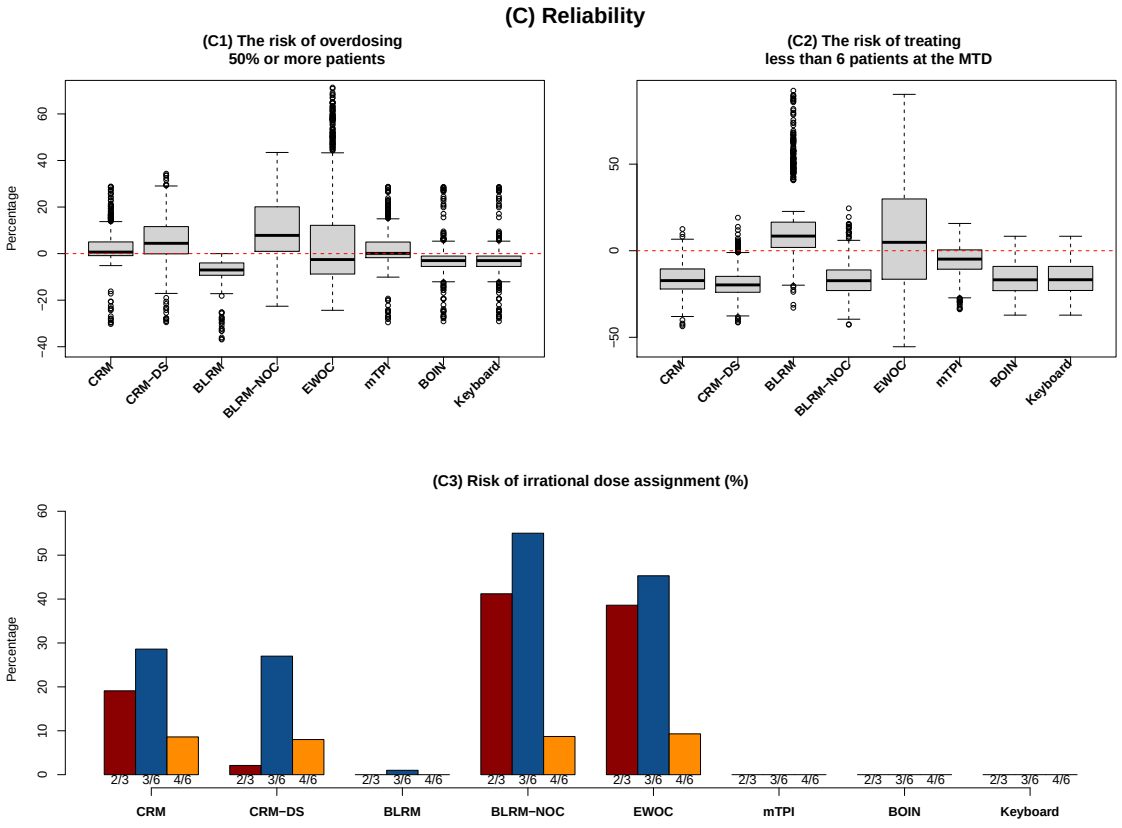

Figure 2.5: Comparison of reliability metrics for the 8 designs with respect to the 3+3 design. C1. Risk of overdosing 50% or more patients; C2. Risk of treating < 6 patients at the MTD; C3. Risk of irrational dose assignment. A smaller value indicates better performance; negative value means that the design outperforms the 3+3 design.

Table 2.1: Escalation and De-escalation rules for the mTPI, BOIN and Keyboard designs under their default settings for a target toxicity rate of $\phi = 0.2$.

	Number of patients treated at the current dose															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
mTPI Design																
Escalate if number of DLTs $\leq$	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
De-escalate if number of DLTs $\geq$	1	2	2	2	3	3	4	4	4	5	5	5	5	6	6	6
BOIN Design																
Escalate if number of DLTs $\leq$	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2	2
De-escalate if number of DLTs $\geq$	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
Keyboard Design																
Escalate if number of DLTs $\leq$	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2
De-escalate if number of DLTs $\geq$	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4

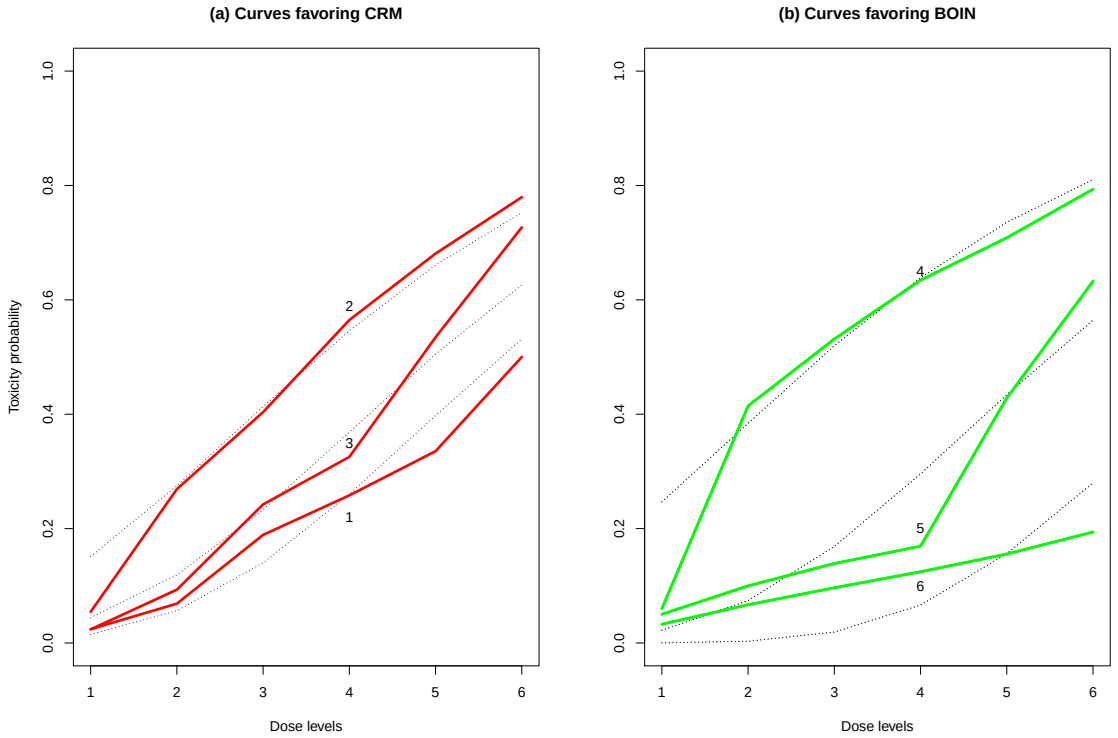

Figure 2.6: Medians of clustered dose-toxicity curves favoring CRM and BOIN designs. Dotted lines are the best model-fitted curves from the CRM design.

CHAPTER 3

BOP2: Bayesian Optimal Design for Phase II Clinical Trials with Simple and Complex Endpoints

3.1 Introduction

This chapter is based on "BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints", published in Statistics in Medicine (2017) [49] coauthored with Jack J. Lee and Ying Yuan. Permission from the journal has been granted for use in conjunction with the dissertation.

Traditionally, phase II clinical oncology trials have focused on binary efficacy endpoints, e.g., tumor response, but they have become much more complicated with the advent of novel molecular targeted agents and immunotherapy. The endpoints for such treatments may be ordinal or multivariate, and the investigators are often interested in simultaneously monitoring multiple types of events in the trial, as illustrated by the following trial examples.

Example 1. Binary efficacy endpoint The aim of a phase II trial is to evaluate the efficacy of pembrolizumab in patients with advanced small bowel adenocarcinomas. The primary endpoint is the ORR, defined using RECIST, version 1.1. The treatment is regarded as futile if the ORR $\leq 20\%$ and promising if the ORR $\geq 40\%$. This example is used to illustrate the standard case with a binary efficacy endpoint.

Example 2. Nested efficacy endpoints The aim of a phase II clinical trial is to assess the efficacy of nivolumab in patients with Hodgkin's lymphoma who have not experienced a successful outcome following an autologous stem cell transplant. The revised International Working Group Criteria for Malignant Lymphoma [50] is used to define the efficacy of treatments for lymphoma, categorized as one of four levels of decreasing desirability: complete remission (CR), defined as the disappearance of all evidence of disease; partial remission (PR), defined as the regression of measurable disease and no new sites; stable disease (SD), defined as failure to attain CR, PR or progressive disease (PD); and PD, defined as evidence of any new lesion or an increase in lesion volume $\geq 50\%$ from the nadir of previously involved sites. In this trial, although both CR and PR are regarded as favorable responses, CR is substantially more desirable. The treatment is regarded as promising if (1) the probability of achieving CR or PR $\geq 30\%$ or (2) the probability of achieving CR $\geq 15\%$, where the endpoint of the second condition is a part of the endpoints of the first condition.

Example 3. Co-primary efficacy endpoints The primary objective of a phase II trial is to evaluate the efficacy of trebananib administered at 15 mg/kg IV per week in patients with persistent or recurrent carcinoma of the endometrium [51]. The trial has two co-primary efficacy endpoints: the objective response rate (ORR) and the event-free survival at 6 months (EFS6). The objective response (OR) is defined using the Response Evaluation Criteria in Solid Tumors (RECIST), version 1.1 [52]. The event-free survival is defined as the length of time from the initiation of the treatment to disease progression, death, or beginning a subsequent therapy. The null hypothesis is that the ORR $\leq 10\%$ and EFS6 $\leq 20\%$. In other words, the treatment is regarded as futile only if the ORR $\leq 10\%$ and EFS6 $\leq 20\%$. Clinically

significant differences are defined as a 20% increase in PFS6, or a 15% increase in ORR.

Example 4. Jointly monitoring efficacy and toxicity In a phase II clinical trial, patients with recurrent indolent non-follicular lymphoma are treated with lenalidomide in combination with rituximab [53]. Lenalidomide is administered at 20 mg/day for days 1-21, and rituximab is administered at 375 mg/m^2 once on day 14 of every 28 days. The primary efficacy endpoint is the ORR as defined using the 1999 Cheson criteria. Because of large uncertainty regarding the safety of the combination treatment, the trial also monitors dose-limiting toxicity, defined according to the National Cancer Institute Common Terminology Criteria for Adverse Events. The lowest acceptable ORR is 45% and the highest acceptable toxicity rate is 30%.

We propose a flexible Bayesian optimal phase II (BOP2) design that is capable of handling the aforementioned trials in a unified framework. We use a Dirichlet-multinomial model to embrace different types of endpoints. At each interim, the go/no-go decision is made by evaluating the posterior probabilities of the events of interest. The BOP2 design explicitly controls the type I error rate and is optimal in the sense that it optimizes power or minimizes the expected sample size under the null hypothesis. Thall and colleagues [54, 55, 56] proposed Bayesian sequential monitoring designs for multiple response outcomes (e.g., toxicity and efficacy). Compared to these designs, the advantage of the proposed BOP2 design includes (1) offering a more flexible framework to monitor multiple events simultaneously, including nested or co-primary endpoints; (2) explicitly controlling the type I error rate, thereby bridging the gap between Bayesian designs and frequentist designs and also rendering the proposed Bayesian design more accessible to a wide range of users and regulatory agencies; and (3) allowing the cutoffs of the stopping rule to vary with

the interim sample size, which improves the power of the design, as demonstrated in our simulation studies.

3.2 Methods

3.2.1 Probability model

Although the endpoints of the aforementioned trials take different forms, they can be unified and represented by a random variable Y that follows a multinomial distribution,

$$Y \sim Multinom(\theta_1, \ldots, \theta_K),$$

where $\theta_k = \Pr(Y = k)$ is the probability that Y belongs to the kth category, $k = 1, \cdots, K$. The K categories can be the actual levels of a single endpoint or the combinational levels of multiple categorical endpoints. For example, in trial example 2, Y is the ordinal outcome, with $Y = 1, 2, 3$ and 4 denoting CR, PR, SD and PD, respectively. In trial example 3, Y is a multinomial variable with four categories where 1 = (OR, EFS6), 2 = (OR, no EFS6), 3 = (no OR, EFS6) and 4 = (no OR, no EFS6). Similarly, in trial example 4, Y is a multinomial variable with four categories: 1 = (toxicity, OR), 2 = (no toxicity, OR), 3 = (toxicity, no OR) and 4 = (no toxicity, no OR). Trial example 1 can be viewed as a special case of trial example 2 by ignoring EFS6, where Y has only two categories (i.e., OR or no OR). In this case, the multinomial distribution degenerates to a binomial distribution.

Suppose that at an interim time, a total of n patients have been enrolled into the trial and their endpoints have been fully evaluated. Let $\mathcal{D}_n = (x_1, \cdots, x_K)$ denote the interim data, and x_k denote the number of patients with $Y = k$, where $\sum_{k=1}^{K} x_k = n$. Assuming that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ follows a Dirichlet prior,

$$(\theta_1, ..., \theta_K) \sim Dir(a_1, \ldots, a_K),$$

where $a_1, \ldots, a_K$ are hyperparameters, the posterior distribution of $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta}|\mathcal{D}_n \sim Dir(a_1 + x_1, \cdots, a_K + x_K).$$

We set $\sum_{k=1}^{K} a_k = 1$ such that the prior is vague and equivalent to a prior sample size of 1. In the special case that Y is a binary outcome (e.g., trial example 1), this Dirichlet-multinomial model becomes a standard beta-binomial model.

3.2.2 Trial design

Let N denote the maximum sample size of the trial. The proposed BOP2 design consists of R interim looks, which occur when the number of enrolled patients reaches $n_1, \cdots, n_R$, and a final look when all N patients are enrolled. At each of these looks, the go/no-go decision is made based on the accumulating data, as described below. In other words, patients are enrolled in $R+1$ cohorts of size n_1, $n_2 - n_1$, ..., $n_R - n_{R-1}$ and $N - n_R$, respectively, and the go/no-go decision is made after each cohort is enrolled. When $R = N - 1$, we obtain a full sequential design in which the go/no-go decision is continuously assessed after each patient. For notational brevity, we suppress the subscript of the interim sample size when this does not cause confusion.

Let $C(n)$ denote a probability cutoff, which is a function of the interim sample size n. Under the proposed design, the go/no-go decision at each interim is made based on the posterior probability of the events of interest. Specifically, for the four trial examples, the interim stopping rule is described as follows. At an interim look, terminate the trial if

(*Example 1*): $\Pr(\theta_1 \leq 0.2|\mathcal{D}_n) > C(n)$

(*Example 2*): $\Pr(\theta_1 \leq 0.15|\mathcal{D}_n) > C(n)$ and $\Pr(\theta_1+\theta_2 \leq 0.3|\mathcal{D}_n) > C(n)$

(*Example 3*): $\Pr(\theta_1 + \theta_2 \leq 0.1|\mathcal{D}_n) > C(n)$ and $\Pr(\theta_1 + \theta_3 \leq 0.2|\mathcal{D}_n) >$

$C(n)$

(*Example 4*): $\Pr(\theta_1 + \theta_2 \leq 0.45|\mathcal{D}_n) > C(n)$ or $\Pr(\theta_1 + \theta_3 > 0.3|\mathcal{D}_n) > C(n)$.

Unlike most existing Bayesian designs [33, 54, 55], which assume a constant cutoff, here we allow the cutoff $C(n)$ to be a function of the interim sample size n. As we show later, such modification is important and substantially improves the power of the design. Although these stopping rules have different clinical interpretations, the go/no-go decisions are all based on the evaluation of a set of the posterior probabilities of the linear combination of the model parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)^T$, for example,

$$\Pr(\boldsymbol{b\theta} \leq \phi|\mathcal{D}_n) > C(n), \tag{3.1}$$

where $\boldsymbol{b}$ is a design vector with elements of 0 and 1, and ϕ is a prespecified threshold. For example, in trial example 2, the stopping rule involves the evaluation of two posterior probabilities, with $\boldsymbol{b} = (1, 0, 0, 0)$ and $\phi = 0.15$, and $\boldsymbol{b} = (1, 1, 0, 0)$ and $\phi = 0.3$, respectively; and in trial example 3, the stopping rule involves the evaluation of two posterior probabilities, with $\boldsymbol{b} = (1, 1, 0, 0)$ and $\phi = 0.1$, and $\boldsymbol{b} = (1, 0, 1, 0)$ and $\phi = 0.2$, respectively.

The evaluation of the posterior probability in (3.1) is facilitated by the following property of the Dirichlet distribution.

Property 1. Given $\theta \sim Dir(a_1 + x_1, \cdots, a_K + x_k)$ and a design vector $\boldsymbol{b} = (b_1, \cdots, b_K)$ with elements of 0 and 1, $\boldsymbol{b}\theta$ follows a Beta distribution $Beta(\sum_{k=1}^K b_k(a_k + x_k), \sum_{k=1}^K (1 - b_k)(a_k + x_k))$.

As a result, $\Pr(\boldsymbol{b\theta} \leq \phi|\mathcal{D}_n)$ can be easily evaluated as

$$\Pr(\boldsymbol{b\theta} \leq \phi|\mathcal{D}_n) = B\big(\phi; \sum_{k=1}^K b_k(a_k + x_k), \sum_{k=1}^K (1 - b_k)(a_k + x_k)\big),$$

where $B(\phi;\, \zeta, \xi)$ is the cumulative distribution function of a Beta distribution with parameters ζ and ξ, evaluated at value ϕ. This property of $\Pr(\boldsymbol{b\theta} \leq \phi | \mathcal{D}_n)$ leads to the following result.

Lemma 1. $\Pr(\boldsymbol{b\theta} \leq \phi | \mathcal{D}_n)$ is a monotonic function of $\sum_{k=1}^{K} b_k x_k$.

The monotonicity of $\Pr(\boldsymbol{b\theta} \leq \phi | \mathcal{D}_n)$ is important in practice because it allows us to enumerate the stopping boundary prior to the onset of the trial, similar to Simon's two-stage design, as shown in Table 3.1. For example, the row "Example 2" shows the interim stopping boundary in terms of the number of patients with CR and CR/PR for our trial example 2. During trial conduct, we do not need to carry out any complicated calculations; rather, we just need to count the number of relevant events and make the go/no-go decision based on whether that count exceeds the boundary. For example, after 20 patients are treated, if the number of CR responses is ≤ 3 and the number of CR/PR responses is ≤ 5, we terminate the trial early. This property makes the BOP2 design very easy to implement in practice.

3.2.3 Optimizing design parameters

Suppose that appropriate null hypothesis H_0 and alternative hypothesis H_1 have been chosen to reflect clinical interests, where H_0 specifies the value of $\boldsymbol{\theta}$, under which the treatment is deemed as futile; and H_1 specifies the value of $\boldsymbol{\theta}$, under which the treatment is deemed as promising. For example, for trial example 2, $H_0 : \theta_1 = 0.15$ and $\theta_1 + \theta_2 = 0.3$, and a reasonable alternative hypothesis is $H_1 : \theta_1 = 0.25$ and $\theta_1 + \theta_2 = 0.5$. With complicated endpoints (e.g., two co-primary endpoints), the specification of H_1 is less straightforward and should be determined through consultation with clinicians to reflect a desirable outcome that is feasible in practice. We reject H_0 and claim that the treatment is promising if the stopping

boundaries are never crossed throughout the trial (including at the end of the trial). The type I error rate and statistical power are defined as the probability of rejecting H_0 under H_0 and H_1, respectively.

The operating characteristics of the BOP2 design rely on the specification of the probability cutoff $C(n)$. Although any reasonably flexible monotonically decreasing function may be used, one particular function of $C(n)$ that is simple and yields good operating characteristics is the following two-parameter power function

$$C(n) = 1 - \lambda(n/N)^{\gamma}, \tag{3.2}$$

where λ and γ are tuning parameters. We require that $\gamma > 0$ such that $C(n)$ is monotonically decreasing with n/N, the fraction of the accumulated information. The rationale is that at the beginning of the trial, data are sparse and a more relaxed stopping rule with a larger value of $C(n)$ may be preferred to avoid terminating the trial accidentally. When the trial proceeds and information accumulates, we have less uncertainty regarding the endpoint of interest and thus it is desirable to have a more stringent stopping rule with a smaller value of $C(n)$ to terminate the trial for an inefficacious treatment. The remaining questions are how to choose the tuning parameters λ, γ and sometimes N in (3.2) to optimize the performance of the design according to a certain criterion.

We first consider how to choose the tuning parameters λ and γ when sample size N is fixed, for example, due to a fixed budget or limited accrual. Our strategy is to choose λ and γ to maximize the power of the BOP2 design, while controlling the type I error rate at a certain prespecified level. This can be done as follows:

Step 1: Elicit from clinicians H_0 and H_1, and the desirable type I error rate.

Step 2: Find the values of (λ, γ) that yield the desirable type I error rate, which can

be done through a grid search.

Step 3: Among the set of (λ, γ) identified in Step 2, select the one that yields the maximum statistical power as the optimal design parameters.

Although the BOP2 design is a Bayesian design, it is still essential to ensure that the design has desirable frequentist operating characteristics (e.g., type I rate and power). In general, a good Bayesian design should demonstrate reasonable frequentist operating characteristics [57]. Explicitly controlling the type I error rate is an important feature that distinguishes the BOP2 design from most existing Bayesian phase II designs. This feature bridges the gap between Bayesian designs and frequentist designs, making the BOP design accessible to a wide range of users and regulatory agencies.

An alternative optimization strategy is to choose λ and γ, as well as sample size N, to minimize the expected sample size under H_0, i.e., $E(N|H_0)$, given prespecified type I and II error rates. This optimization criteria was used in Simon's optimal design. In this approach, N is not fixed, but a design parameter to be optimized. The procedure to determine the values of (λ, γ, N) that minimize $E(N|H_0)$ can be described as follows:

Step 1: Elicit from clinicians H_0 and H_1, and desirable type I and II error rates.

Step 2: Find the values of (N, λ, γ) that yield the desirable type I and II error rates, which can be done through a grid search.

Step 3: Among the set of (N, λ, γ) identified in Step 2, select the one that yields the smallest $E(N|H_0)$ as the optimal design parameters.

In Step 2, we have two constraints (i.e., type I and II error rates), but need to determine the values of three unknown parameters (N, λ, γ). Thus, in principle, there

are an infinite number of possible solutions. We circumvent this issue by restricting the value of N within the range of (N_{min}, N_{max}), where N_{max} is the maximum sample size that we can afford in practice, which is often determined by budget, accrual rate or other practical factors. N_{min} is the minimal sample size for the trial, which has little impact on the operating characteristics of the design as long as it is reasonably small, such as $N_{min} = 10$. Given a specific value of N, we can uniquely determine the values of λ and γ based on the two constraints through a grid search. One potential limitation of this optimization strategy is that we do not have a direct control on sample size N, and the value of N that minimizes $E(N|H_0)$ may be excessively large for practical use in some cases. When this is a concern, the minimax criterion can be used to optimize the design. That is, instead of minimizing $E(N|H_0)$, we choose (λ, γ, N) to minimize the maximum sample size.

3.3 Web application

To facilitate the use of the BOP2 design, we develop an easy-to-use web application using Shiny. Figure 3.2 shows the graphical user interface of the application. After users input their design parameters (e.g., the maximum sample size, cohort size, desirable type I and II errors and the type of endpoints to be monitored), the web application generates the operating characteristics and stopping boundary of the BOP2 design that can be included in the trial protocol. As described previously, similar to Simon's two-stage design, one important advantage of the BOP2 design is that its stopping boundary can be enumerated and included in the trial protocol prior to the onset of the trial. When conducting the trial, we simply count the number of relevant events and make the go/no-go decision based on whether that count exceeds the boundary. Our web application will be freely available at

`http://www.trialdesign.org` and also the author's website.

3.4 Simulation study

In the following simulation studies, we controlled the type I error rate at 0.1 for all designs. Due to the discrete nature of the observed data, a cutoff that yields a type I error rate of exactly 0.1 does not usually exist. In these cases, we chose the cutoff that yielded a type I error rate closest to and not higher than the nominal value. For clarity, in what follows, we focus on the BOP2 design that maximizes power. Interim analyses were carried out after the first 10 patients were treated, then after every 5 additional patients were treated, with the maximum sample size $N = 40$. The results of the BOP2 design that minimizes $E(N|H_0)$ are similar and provided in the Appendix.

To evaluate the performance of the designs, we considered the following three metrics. (1) The percentage of rejecting the null hypothesis (PRN) is defined as the percentage of the simulated trials in which H_0 is rejected. The PRN is the type I error rate (or power) when H_0 is (or is not) true. The PRN can be also explained as the percentage of claiming that the new treatment is effective. (2) The percentage of early termination (PET) is defined as the percentage of trials that are terminated early. (3) The actual sample size is defined as the average sample size actually used in 10,000 simulated trials.

3.4.1 Binary efficacy endpoint

We first evaluated the operating characteristics of the BOP2 design under the conventional setting with a simple binary efficacy outcome (i.e., OR/no OR), as illustrated by trial example 1. We selected four pairs of H_0 and H_1, and compared the BOP2 design to the Bayesian design proposed by Thall and Simon [33], denoted

as the TS design. In the first scenario, for example, the TS design employed a beta-binomial model and terminated the trial if $\Pr(\theta_1 < 0.2|\mathcal{D}_n) > C$, where C is a fixed cutoff. We calibrated the value of C to control the type I error rate of the TS design at 0.1, which matches that of the BOP2 design. Table 3.1 provides the stopping boundaries of the BOP2 design in terms of the observed number of patients who had responses, which were used to make the go/no-go decision at each interim look for scenario 1 in Table 3.2.

Table 3.2 shows the performance of the two designs under four different pairs of H_0 and H_1 on the ORR, represented as scenarios 1-4. In general, the BOP2 design yielded substantially higher power than the TS design. For example, in scenario 1, where the null ORR is 0.2 and the alternative ORR is 0.4, when the true ORR is 0.4, the power of the BOP2 design is 88.3%, whereas the power of the TS design is only 76.4%. In addition, compared to the TS design, the BOP2 design had a lower risk of incorrectly terminating the trial when the treatment is actually effective. For example, when the true ORR is 0.4, the TS design incorrectly terminated the trial 23.5% of the time, while the BOP2 design incorrectly terminated the trial 11.4% of the time. Under the null hypothesis, the TS design had a higher probability of terminating the trial than the BOP2 design. Because the TS design had a high tendency of terminating the trial, it has smaller (actual) sample sizes than the BOP2 design. Due to the discrete nature of the observed data and different ways of defining the cutoff (i.e., the TS design uses a fixed cutoff C and the BOP2 design uses an adaptive cutoff $C(n)$), in some cases, it is not possible to exactly match the type I error rate of two designs to 0.1. That is why in some scenarios (e.g., scenario 1), the type I error of the TS design is slightly lower than the BOP2 design, but slightly higher in other scenarios (e.g., scenario 2).

Figure 3.1 contrasts the stopping boundaries of the two designs under scenario 1 in Table 3.2. At each interim look, the trial is terminated if the number of observed responses does not exceed the line in the plot. In the early stages of the trial, the BOP2 design has lower boundaries because of the relaxed stopping rules. At later stages, the BOP2 design applies more stringent stopping rules to yield higher boundaries than the TS design. Similar patterns are observed in the boundary plots for other scenarios in Table 3.2.

3.4.2 Nested efficacy endpoints

Table 3.3 shows the simulation results for the ordinal endpoint under the setting of trial example 2, and Table 3.1 shows the corresponding boundaries. Interim monitoring started after the first 10 patients were enrolled, and then was performed after every 5 patients was enrolled. The null hypothesis is scenario 1, i.e., H_0 : $\Pr(\text{CR}) = 0.15$ and $\Pr(\text{CR/PR}) = 0.3$, and the alternative hypothesis is scenario 7, i.e., H_1: $\Pr(\text{CR}) = 0.25$ and $\Pr(\text{CR/PR}) = 0.5$. We compared the BOP2 design with the TS design, which regards CR/PR as response and SD/PD as nonresponse, as is often done in practice. Thus, the TS design employed the beta-binomial model and terminated the trial if $\Pr(\theta_1 < 0.3|\mathcal{D}_n) > C$, where C is a fixed cutoff. We calibrated the value of C to control the type I error rate of the TS design at 0.1.

As shown in Table 3.3, the BOP2 design generally has more power than the TS design. For example, in scenario 7, where $\Pr(\text{CR}) = 0.25$ and $\Pr(\text{CR/PR}) = 0.5$, the power of the BOP2 design is 85.5%, whereas that of the TS design is 74.2%. Comparing the first two scenarios, we observe that the BOP2 design can increase the PRN from 8.7% to 24.2% because the true CR rate increases from 0.15 to 0.20. This exactly fulfills our expectation of the BOP2 design, which can monitor nested endpoints simultaneously. In contrast, the TS design could not distinguish these

two scenarios because the CR/PR rates are both 0.30. In addition, the TS design tended to incorrectly terminate the trial more frequently than the BOP2 design when the treatment actually was effective. For example, in scenario 7, the TS design terminated the trial early 25.7% of the time, whereas the BOP2 design terminated the trial early 9.9% of the time. Again, due to a high tendency of terminating the trial early, the actual sample size of the TS design is smaller than that of the BOP2 design.

3.4.3 Co-primary efficacy endpoints

Table 3.4 shows the simulation results under the setting of trial example 3, with two co-primary efficacy endpoints (i.e., ORR and EFS6), and Table 3.1 shows the corresponding boundaries. The H_0 and H_1 are scenarios 1 and 7, respectively. We compared the BOP2 design to a Bayesian design inspired by the method of Thall, Simon and Estey [54], and denoted the latter as the TSE design. For fair comparison, the TSE design used the same model and stopping rule as the BOP2 design, except that a fixed cutoff C was used in the stopping rule, as suggested by Thall, Simon and Estey [54]. The results are generally similar to those described previously. That is, the BOP2 design yielded higher power and was less likely to incorrectly terminate the trial than the TSE design.

3.4.4 Efficacy and toxicity endpoints

Table 3.5 shows the simulation results under the setting of trial example 4, where we simultaneously monitored efficacy and toxicity. Scenarios 1 and 7 show H_0 and H_1, respectively. We compared the BOP2 design with the TSE design, which used the same model and stopping rule as the BOP2 design, except that a fixed cutoff C was used in the stopping rule. Again, given the same type I error rate (i.e., PRN

in scenario 1), the BOP2 design outperformed the fixed-cutoff design with higher power and smaller risk of incorrectly terminating the trial.

The corresponding boundaries are shown in Table 3.1. The use of stopping boundaries in this case is slightly different from that in the previous cases. For example, after 30 patients are treated, the BOP2 design terminates the trial if either the number of responses is ≤ 13 or the number of toxicities is ≥ 10.

Table 3.1: Stopping boundaries of the BOP2 design for four trial examples. Maximum sample size is 40.

			Number of patients treated						
Trial	Stop the trial if		10	15	20	25	30	35	40
Example 1		# of ORR $\leq$	1	2	4	5	7	9	10
Example 2	and	# of CR $\leq$	0	1	3	4	5	7	9
		# of CR/PR $\leq$	2	3	5	8	10	13	16
Example 3	and	# of ORR $\leq$	0	1	2	3	4	5	7
		# of EFS $\leq$	1	2	4	5	7	9	12
Example 4	or	# of Responses $\leq$	2	5	7	10	13	16	19
		# of Toxicities $\geq$	5	6	8	9	10	11	12

Table 3.2: Percentage of rejecting the null (PRN), percentage of early termination (PET), and actual sample size under the BOP2 design and TS design (Thall and Simon, 1994) with a binary endpoint as described in trial example 1.

	Response rate	PRN(%)		PET(%)		Sample size	
Scenario	(ORR)	BOP2	TS	BOP2	TS	BOP2	TS
1	0.20[§]	9.6	9.4	88.8	89.8	20.2	15.3
	0.30	55.2	42.6	46.2	56.7	31.0	24.9
	0.40[†]	88.3	76.4	11.4	23.5	37.6	33.6
	0.50	98.2	93.3	1.8	6.7	39.5	38.1
2	0.30[§]	9.4	10.0	82.8	89.5	22.2	15.6
	0.40	48.2	40.3	41.4	59.4	31.9	24.3
	0.50[†]	86.7	74.0	10.3	25.9	37.7	32.9
	0.60	98.9	92.7	1.8	7.3	39.5	37.9
3	0.40[§]	10.0	10.0	84.3	89.0	21.6	15.4
	0.50	47.5	38.3	46.0	61.1	31.0	23.5
	0.60[†]	86.3	72.1	11.9	27.9	37.5	32.4
	0.70	98.2	92.5	1.7	7.5	39.6	37.9
4	0.50[§]	7.2	7.2	79.5	92.4	24.7	13.4
	0.60	43.0	29.6	36.3	70.2	33.9	20.0
	0.70[†]	87.3	61.9	6.4	38.1	38.9	28.9
	0.80	99.6	87.6	0.2	12.4	40.0	36.3

[§]: null hypothesis; [†]: alternative hypothesis.

Table 3.3: Percentage of rejecting the null (PRN), percentage of early termination (PET), and the actual sample size under the BOP2 design and TS design (Thall and Simon, 1994) with nested endpoints as described in trial example 2.

			PRN(%)		PET(%)		Sample size	
Scenario	$\boldsymbol{\theta}$	(CR, CR/PR)	BOP2	TS	BOP2	TS	BOP2	TS
1	(0.15,0.15,0.30,0.40)[§]	(0.15, 0.30)	8.7	9.9	82.1	89.6	25.4	15.7
2	(0.20,0.10,0.30,0.40)	(0.20, 0.30)	24.2	9.6	63.8	89.9	29.0	15.6
3	(0.20,0.15,0.30,0.35)	(0.20, 0.35)	30.6	22.9	56.3	76.6	31.0	19.7
4	(0.20,0.20,0.30,0.30)	(0.20, 0.40)	45.9	40.7	41.0	59.2	33.8	24.4
5	(0.20,0.25,0.30,0.25)	(0.20, 0.45)	66.3	58.7	24.1	41.2	36.4	29.0
6	(0.25,0.20,0.30,0.25)	(0.25, 0.45)	72.3	59.0	19.3	40.9	37.1	29.1
7	(0.25,0.25,0.25,0.25)[†]	(0.25, 0.50)	85.5	74.2	9.9	25.7	38.5	33.0
8	(0.30,0.25,0.25,0.20)	(0.30, 0.55)	95.7	85.2	3.0	14.8	39.5	35.9

[§]: null hypothesis; [†]: alternative hypothesis.

Table 3.4: Percentage of rejecting the null (PRN), percentage of early termination (PET), and the actual sample size under the BOP2 design and TSE design (Thall, Simon and Estey, 1995) with two co-primary efficacy endpoints as described in trial example 3.

			PRN(%)		PET(%)		Sample size	
Scenario	$\boldsymbol{\theta}$	(ORR, EFS)	BOP2	TSE	BOP2	TSE	BOP2	TSE
1	(0.05,0.05,0.15,0.75)[§]	(0.10, 0.20)	7.2	7.3	80.9	92.3	24.5	13.7
2	(0.05,0.10,0.15,0.70)	(0.15, 0.20)	23.9	17.4	58.9	82.3	29.7	16.5
3	(0.10,0.10,0.15,0.65)	(0.20, 0.25)	56.5	37.1	29.7	62.8	35.0	22.0
4	(0.10,0.15,0.15,0.60)	(0.25, 0.25)	79.7	53.0	12.8	47.0	37.8	26.3
5	(0.10,0.15,0.20,0.55)	(0.25, 0.30)	85.9	60.7	7.6	39.3	38.7	28.5
6	(0.15,0.15,0.10,0.60)	(0.30, 0.25)	91.3	64.4	6.6	35.6	38.6	29.5
7	(0.15,0.15,0.20,0.50)[†]	(0.30, 0.35)	96.1	75.5	2.4	24.5	39.5	32.8
8	(0.15,0.15,0.25,0.45)	(0.30, 0.40)	98.5	82.6	0.8	17.4	39.8	34.8

[§]: null hypothesis; [†]: alternative hypothesis.

Table 3.5: Percentage of rejecting the null (PRN), percentage of early termination (PET), and the actual sample size under the BOP2 design and the TSE design (Thall, Simon and Estey, 1995) with jointly monitoring efficacy and toxicity endpoints as described in trial example 4.

			PRN(%)		PET(%)		Sample size	
Scenario	$\boldsymbol{\theta}$	(Eff, Tox)	BOP2	TSE	BOP2	TSE	BOP2	TSE
1	(0.15,0.30,0.15,0.40)[§]	(0.45, 0.30)	9.3	9.2	85.3	89.6	22.1	15.2
2	(0.20,0.30,0.15,0.35)	(0.50, 0.35)	7.5	7.1	88.5	91.3	20.6	14.6
3	(0.10,0.30,0.15,0.45)	(0.40, 0.25)	6.7	7.5	88.7	91.4	21.2	14.7
4	(0.15,0.35,0.10,0.40)	(0.50, 0.25)	30.7	25.5	61.7	73.3	28.2	20.1
5	(0.15,0.35,0.05,0.45)	(0.50, 0.20)	41.0	33.8	51.2	65.3	30.7	22.4
6	(0.15,0.40,0.05,0.40)	(0.55, 0.20)	60.8	48.6	33.5	50.8	33.9	26.1
7	(0.18,0.42,0.02,0.38)[†]	(0.60, 0.20)	74.6	59.8	22.0	39.7	36.0	29.1
8	(0.15,0.50,0.05,0.30)	(0.65, 0.20)	82.6	69.4	15.1	30.3	37.0	31.7

[§]: null hypothesis; [†]: alternative hypothesis.

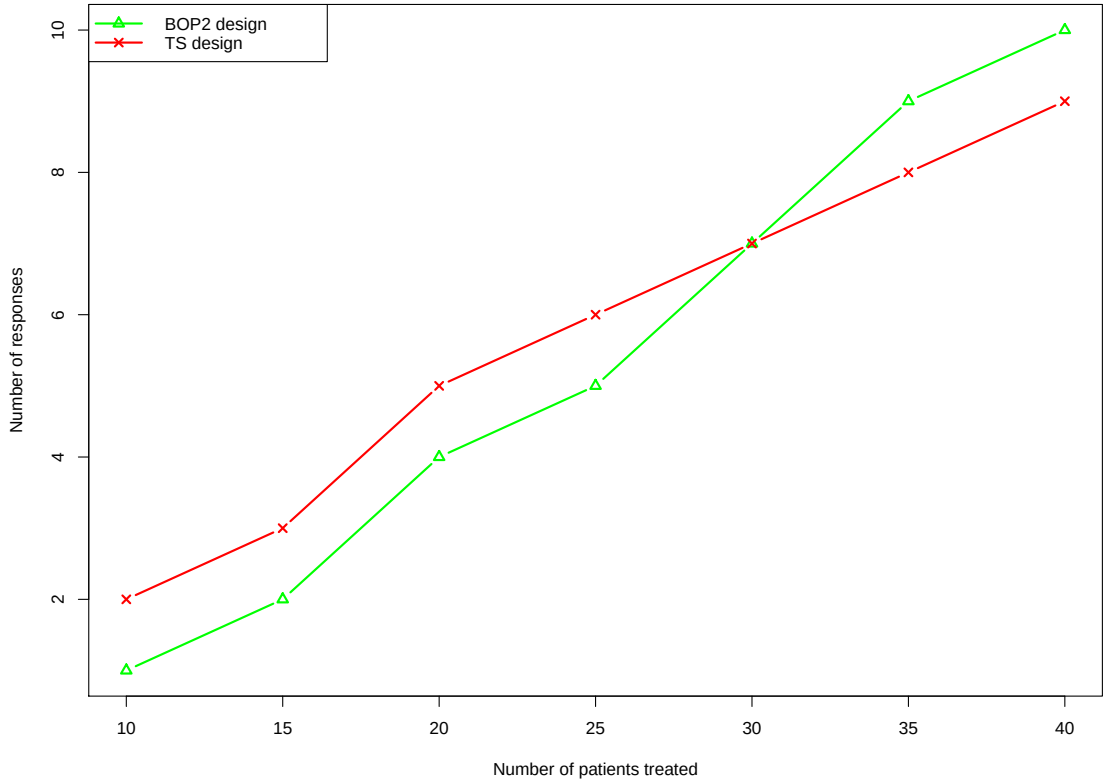

Figure 3.1: Stopping boundaries of BOP2 design and TS design for the binary efficacy endpoint (i.e., trial example 1) under scenario 1 shown in Table 3.2. The maximum sample sizes of the two designs are 40.

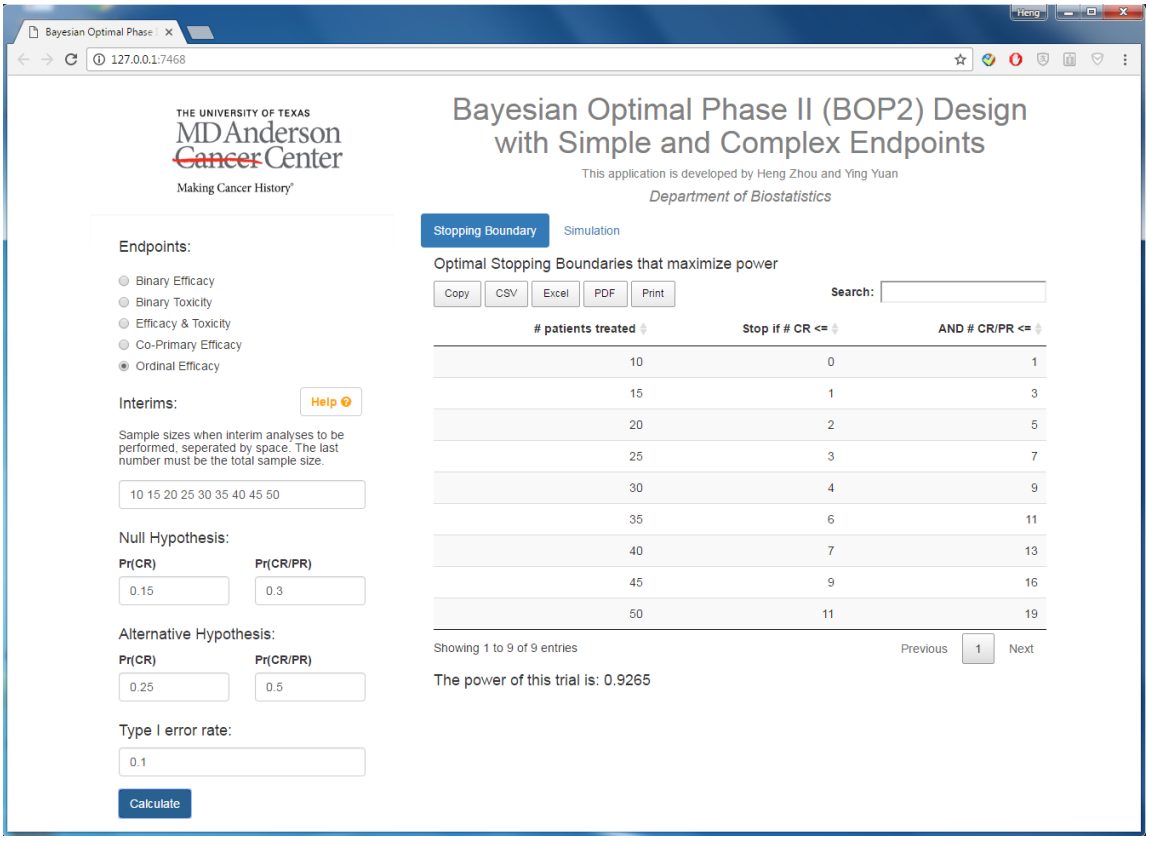

Figure 3.2: Web application for the BOP2 design

CHAPTER 4

Bayesian Optimal Phase II Design for Cancer Immunotherapy

4.1 Introduction

The objective response rate is a short-term endpoint in phase II clinical trials, which can be assessed usually in the first cycle of treatment. The long-term survival and durable response rates are also desirable endpoints, which indicate the patients being cured. We propose an optimal phase II clinical trial design to simultaneously model the objective response rate (ORR) and the cure rate (e.g., long-term survival and durable tumor response). Consider the trial example 2 described in section 3.1. Suppose the new treatment is expected to achieve more than 10% in durable clinical response rate, the interim decision should be based on the rule: the treatment is regarded as promising if (1) the probability of achieving CR or PR rate $\geq 30\%$, or (2) the probability of achieving CR rate $\geq 15\%$, or (3) the probability of achieving cure rate $\geq 10\%$. In order to model the cure rate among the patients population, we use the mixture cure rate model proposed by Berkson and Gage (1952) since it could capture the information of both "cured" and "non-cured" subsets at the same time [58]. Following the BOP2 design, we use the non-constant posterior probability cutoff to make interim decision and optimize the design parameters while controlling the type I error rate. Therefore the proposed design is an extension to the BOP2

design, which we call BOP2-C design.

4.2 Methods

4.2.1 Probability model

In the first cycle of treatment, we define Y as the ordinal outcome, with $Y = 1, 2, 3$ and 4 denoting CR, PR, SD and PD, respectively. Thus Y follows the multinomial distribution such that

$$Y \sim Multinom(\pi_1, \pi_2, \pi_3, \pi_4),$$

where $\pi_k = \Pr(Y = k), k = 1, 2, 3, 4$, and $\sum_{k=1}^{4} \pi_k = 1$. Suppose that at an interim time, a total of n patients have been enrolled into the trial and their responses have been fully observed. Let $\mathcal{D}_n = (n_1, n_2, n_3, n_4)$ denote the interim data, and n_k denote the number of patients with $Y = k$, where $\sum_{k=1}^{4} n_k = n$. Assuming that $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)^T$ follows a Dirichlet prior,

$$(\pi_1, \pi_2, \pi_3, \pi_4) \sim Dir(a_1, a_2, a_3, a_4),$$

where a_1, a_2, a_3, a_4 are hyperparameters, the posterior distribution of $\boldsymbol{\pi}$ is given by

$$\boldsymbol{\pi}|\mathcal{D}_n \sim Dir(a_1 + n_1, a_2 + n_3, a_3 + n_3, a_4 + n_4).$$

Let θ denote the cure rate in patients population. We model the time to disease progression t with the survival function such that

$$S^*(t) = \theta + (1 - \theta)S(t),$$

where $S(t)$ is the survival function for the non-cured sub-population. We assume Weibull distribution for the time to disease progression in the non-cured patients such that

$$\begin{aligned} S(t) &= \exp\{-(\lambda t)^\alpha\} \\ f(t) &= \alpha\lambda(\lambda t)^{\alpha-1} \exp\{-(\lambda t)^\alpha\} \end{aligned} \tag{4.1}$$

where $f(t)$ is the density function of t. At an interim look, the disease progression of each patient may not be observed. Let δ_i denote the censoring status of the i^{th} patient at the observation time t_i, where $\delta_i = 0$ indicates censored observation. Then the likelihood function of the i^{th} patient with the event time t_i is

$$[(1-\theta)f(t_i)]^{\delta_i}\,[\theta + (1-\theta)S(t_i)]^{1-\delta_i}\,,$$

where $f(t_i)$ and $S(t_i)$ are given by equation 4.1.

4.2.2 Prior specification and posterior estimation

For the Dirichlet prior of $\boldsymbol{\pi}$, we set $\sum_{k=1}^{4} a_k = 1$ such that the prior is vague and equivalent to a prior effective sample size of 1. Since the response of PD in the first cycle of treatment indicates failure in curing patients, the cure rate θ is no larger than $\pi_1 + \pi_2 + \pi_3$. Thus, we assign a uniform prior to θ such that

$$\theta \sim U(0, a_1 + a_2 + a_3).$$

We assign the Gamma priors to the parameters of Weibull distribution (equation 4.1) α and λ as follows,

$$\begin{aligned} \alpha &\sim Ga(a_\alpha, b_\alpha) \\ \lambda &\sim Ga(a_\lambda, b_\lambda) \end{aligned}$$

We set the shape parameters as $a_\alpha = 0.1$, $a_\lambda = 0.05$, and rate parameters as $b_\alpha = b_\lambda = 0.1$, such that the priors are vague.

At each interim look, we update the posterior estimates of $\boldsymbol{\pi}$, θ, α and λ, using the Adaptive Rejection Metropolis Sampling within Gibbs Sampling (ARMS) algorithm [59]. Suppose at the interim n patients have been enrolled, and the number of patients with responses CR, PR, SD and PD are n_1, n_2, n_3, and n_4, respectively. At each iteration, we conduct following steps:

1. Update $\boldsymbol{\pi}|\cdot \sim Dirichlet(a_1+n_1,\ldots,a_4+n_4)$, with the truncation that $\pi_1+\pi_2+\pi_3 \geq \theta$.
2. Update $\theta|\cdot \propto \prod_{i=1}^{n}\left[(1-\theta)f(t_i|\alpha,\lambda)\right]^{\delta_i}\left[\theta+(1-\theta)S(t_i|\alpha,\lambda)\right]^{1-\delta_i}$, with the truncation that $\theta \leq \pi_1+\pi_2+\pi_3$.
3. Update $\alpha|\cdot \propto \prod_{i=1}^{n}\left[(1-\theta)f(t_i|\alpha,\lambda)\right]^{\delta_i}\left[\theta+(1-\theta)S(t_i|\alpha,\lambda)\right]^{1-\delta_i}\times\alpha^{a_\alpha-1}\exp(-b_\alpha\alpha)$.
4. Update $\lambda|\cdot \propto \prod_{i=1}^{n}\left[(1-\theta)f(t_i|\alpha,\lambda)\right]^{\delta_i}\left[\theta+(1-\theta)S(t_i|\alpha,\lambda)\right]^{1-\delta_i}\times\lambda^{a_\lambda-1}\exp(-b_\lambda\lambda)$.

4.2.3 Trial Design

Same as the BOP2 design, The proposed BOP2-C design consists of R interim looks, which occur when the number of enrolled patients reaches $n_1,\cdots,n_R$, and a final look when all N patients are enrolled. That is to say, the patients are enrolled in $R+1$ cohorts of size n_1, n_2-n_1, ..., n_R-n_{R-1} and $N-n_R$, respectively, and the go/no-go decision is made after each cohort is enrolled. For notational brevity, we suppress the subscript of the interim sample size, using n to denote each interim sample size globally. Let $C(n)$ denote a probability cutoff, which is a function of the interim sample size n. Under the BOP2-C design, the go/no-go decision at each interim is made based on the posterior estimation of the CR, CR/PR and cure rates, respectively. Specifically, the proposed design would terminate the trial at an interim look if:

$$\begin{aligned}
\Pr(\pi_1 \leq \phi_1|\mathcal{D}_n) &> C(n) \quad \text{and} \\
\Pr(\pi_1+\pi_2 \leq \phi_2|\mathcal{D}_n) &> C(n) \quad \text{and} \\
\Pr(\theta \leq \phi_3|\mathcal{D}_n) &> C(n).
\end{aligned}$$

ϕ_1,ϕ_2,ϕ_3 are pre-specified thresholds which are usually elicited from clinicians to reflect the null hypothesis H_0 that the treatment is futile. The clinicians also need

to specify a alternative hypothesis H_1 under which the treatment is regarded as promising. An example of the null/alternative hypotheses state that:

$$H_0: \quad \pi_1 = 0.15, \quad \pi_1 + \pi_2 = 0.30, \quad \theta = 0.1$$

$$H_1: \quad \pi_1 = 0.25, \quad \pi_1 + \pi_2 = 0.50, \quad \theta = 0.2$$

The type I error rate and statistical power are defined as the probability of rejecting H_0 under H_0 and H_1, respectively.

We use the same function for $C(n)$ as the BOP2 design such that

$$C(n) = 1 - \lambda_c (n/N)^{\gamma}, \tag{4.2}$$

where λ_c and γ are tuning parameters. As described in section 3.2.3, to select optimized tuning parameters, we can either (1) maximize the power while controlling the type I err rate given the fixed maximum sample size N, or (2) minimize the expected sample size under H_0 while controlling type I and II error rates given unfixed maximum sample size.

4.3 Simulation study

4.3.1 Operating characteristics

We conducted numeric studies to evaluate the operating characteristics of the proposed design. We specify the maximum sample size $N = 120$, and make go/no-go decisions at the interims $n = 40$ and 80. In other words, we enroll three cohorts of patients with cohort size 40, and make interim decisions after each cohort. We assume the arrival time of the patients follows Poisson process, and one cohort of patients, i.e., 40 patients are enrolled within 6 months. We assume the first cycle of treatment is 1 month. That is to say, the tumor response can be fully observed within 1 month for each patient. We make the interim go/no-go decision after the response result

of the last patient in cohort is observed, i.e., 1 month after the patient is enrolled into the trial. We assume Weibull distribution for the time to disease progression in non-cured patients. At the time 1 month, the observation of PD indicates disease progression. Thus we have the survival function for the time to disease progression in non-cured patients at 1 month is

$$S(1) = 1 - \frac{\pi_4}{1-\theta},$$

where π_4 and θ are the true PD rate and cure rate, respectively. Also we assume that 95% of the non-cured patients will experience disease progression within 6 month, i.e., $S(6) = 0.05$. Therefore, we can determine the pair of Weibull parameters for each scenario to simulate the time to disease progression of each patient.

To evaluate the performance of the designs, we considered the following three metrics. (1) The percentage of rejecting the null hypothesis (PRN) is defined as the percentage of the simulated trials in which H_0 is rejected. The PRN is actually the empirical type I error rate (or power) when H_0 is (or is not) true. The PRN can be also explained as the percentage of claiming that the new treatment is promising. (2) The percentage of early termination (PET) is defined as the percentage of trials that are terminated early. (3) The actual sample size is defined as the average sample size across 10,000 simulated trials.

Table 4.1 shows the simulation results. The null hypothesis is scenario 1, i.e., $H_0 : \pi_1 = 0.15$ and $\pi_1 + \pi_2 = 0.35$ and $\theta = 0.10$, and the alternative hypothesis is scenario 5, i.e., $H_1 : \pi_1 = 0.20$ and $\pi_1 + \pi_2 = 0.25$ and $\theta = 0.15$. For the proposed BOP2-C design, We controlled the type I error rate at 0.1, and maximizes the power to get the optimized design parameters. We compared the BOP2-C design with the Thall and Simon's design (TS design) [33], which only monitors the posterior probability of CR/PR response rate, as is often done in practice. Thus, the TS

design employed the beta-binomial model and terminated the trial if $\Pr(\pi_1 + \pi_2 \leq 0.35|\mathcal{D}_n) > C$, where C is a fixed cutoff. We calibrated the value of C to control the type I error rate of the TS design at 0.1.

As shown in Table 4.1, the BOP2-C design generally has more power than the TS design. For example, under the alternative hypothesis scenario 5, the power of the BOP2-C design is 90.1%, compared to 64.0% for the TS design. For scenario 2, the only difference with scenario 1 is that we reversed the values of CR and PR rate. Under this scenario, the PRN of the BOP2-C design should increase because the CR rate is greater than 0.15 while other two remain the same. As expected, the PRN increases to 46.4%, and the trial is slightly less likely to stop early. For the TS design, however, it cannot distinguish scenarios 1 and 2, because the CR/PR rate remains the same. For scenario 3, although the CR rate is less than 0.15 and CR/PR rate is less than 0.35, the cure rate is higher than 0.1. Thus, our design still yields higher power (PRN = 59.1%), while the TS design almost never claims the new treatment is promising (PRN = 0.7%). The comparison among the first three scenarios indicate that the BOP2-C design actually meets our expectation to monitor CR, CR/PR and cure rates simultaneously.

4.3.2 Sensitivity analysis

We also conducted the sensitivity analysis to evaluate the performance of BOP2-C design when the time to disease progression in the non-cured patients does not follow the Weibull distribution. Table 4.2 shows the simulation results when the time to disease progression follows the log-logistic distribution, in contrast to the results shown in Table 4.1. We can see under different assumptions of the time to disease progression, the results are similar to each other in most of the scenarios, especially in terms of the power of design, which indicates that our proposed design

is robust.

Table 4.1: Percentage of rejecting the null (PRN), percentage of early termination (PET), and the actual sample size under the BOP2-C design and TS design (Thall and Simon, 1994). Interim sample sizes are 40, 80, 120.

			PRN(%)		PET(%)		Sample size	
Scenario	$\boldsymbol{\pi}$	θ	BOP2-C	TS	BOP2-C	TS	BOP2-C	TS
1	(0.15,0.20,0.30,0.35)[§]	0.10	9.4	9.4	63.7	88.1	72.0	52.9
2	(0.20,0.15,0.30,0.35)	0.10	46.4	9.2	17.5	88.3	110.6	52.7
3	(0.10,0.20,0.30,0.40)	0.15	59.1	0.7	7.8	98.3	115.6	43.2
4	(0.15,0.25,0.30,0.30)	0.15	70.9	34.4	3.7	63.6	118.0	72.1
5	(0.20,0.25,0.30,0.25)[†]	0.15	90.1	64.0	1.4	35.5	119.2	92.9
6	(0.20,0.25,0.30,0.25)	0.20	98.5	63.9	0.3	35.6	119.9	92.9
7	(0.25,0.20,0.30,0.25)	0.15	95.7	64.7	0.7	64.7	119.6	93.5

[§]: null hypothesis; [†]: alternative hypothesis.

Table 4.2: Percentage of rejecting the null (PRN), percentage of early termination (PET), and the actual sample size of the BOP2-C design under Weibull and Log-logistic assumptions of time to disease progression.

			PRN(%)		PET(%)		Sample size	
Scenario	$\boldsymbol{\pi}$	θ	Weib	Log-L	Weib	Log-L	Weib	Log-L
1	(0.15,0.20,0.30,0.35)[§]	0.10	9.4	9.8	63.7	60.4	72.0	78.5
2	(0.20,0.15,0.30,0.35)	0.10	46.4	32.4	17.5	42.8	110.6	92.2
3	(0.10,0.20,0.30,0.40)	0.15	59.1	55.6	7.8	20.1	115.6	106.9
4	(0.15,0.25,0.30,0.30)	0.15	70.9	66.4	3.7	11.2	118.0	113.1
5	(0.20,0.25,0.30,0.25)[†]	0.15	90.1	85.0	1.4	6.2	119.2	116.2
6	(0.20,0.25,0.30,0.25)	0.20	98.5	96.9	0.3	1.7	119.9	118.8
7	(0.25,0.20,0.30,0.25)	0.15	95.7	92.0	0.7	4.1	119.6	117.2

[§]: null hypothesis; [†]: alternative hypothesis.

CHAPTER 5

Conclusion and Future Work

In chapter 2, we evaluated the operating characteristics of some novel Bayesian phase I trial designs in terms of accuracy, overdose control and reliability. Compared to the 3+3 design, most of these novel designs yield better accuracy for identifying the MTD and allocate more patients to the MTD. Overall, CRM performs well in most metrics. Allowing dose skipping slightly improves the accuracy for identifying the MTD and the allocation of patients to the MTD, but at the cost of substantially increasing the number of overdosed patients and decreasing the design reliability (i.e., a higher risk of overdosing a large percentage of patients). Thus, dose skipping in CRM is generally not recommended, and we suggest restricting dose escalation and de-escalation to one dose level at a time. The performance of BLRM is mixed. BLRM (with the overdose control rule) is excessively conservative and has poor accuracy to identify the MTD and allocate patients to the MTD. Removing the overdose control rule (i.e., BLRM-NOC) improves the accuracy to identify the MTD and allocate patients to the MTD, but at the cost of substantially reduced safety (i.e., treating a high percentage of patients above the MTD) and reliability (i.e., high risk of overdosing a large percentage of patients). The overdose control rule commonly used in BLRM seems to be too conservative, and a more appropri-

ate overdose control rule may be needed to make BLRM work appropriately. The EWOC appears overly conservative. It is safe, but has poor accuracy to identify the MTD. The EWOC has similar average performance as the BLRM, but has larger variation. The model-assisted designs BOIN and keyboard yield good performance that is generally comparable to that of the CRM in terms of accuracy and safety, while often providing smaller variation and better reliability. The mTPI performs well in identifying the MTD and allocating patients to the MTD when the target DLT probability is 0.25, but has lower reliability with a higher risk of overdosing a large percentage of patients and poor allocation of patients to the MTD. The mTPI has a relatively low accuracy to identify the MTD when the target DLT probability is 0.2. Given that BOIN and keyboard are more transparent and simple to implement, they provide attractive approaches to designing phase I clinical trials. The BOIN and keyboard designs have virtually the same performance in every metric. As the BOIN uses the observed DLT rate to determine dose escalation and de-escalation, it is more transparent and assessable for non-statisticians, and is easier to calibrate to fit the design goal. In addition, as noted by a referee, the BOIN has both Bayesian and frequentist interpretations. Its decision rule is equivalent to using the likelihood ratio test to determine dose escalation/de-escalation [19], making it appealing to wider audiences. In contrast, the mTPI/mTPI2 and keyboard designs only have a Bayesian interpretation and require specification of the prior and calculation of the posterior distribution.

In our Monto Carlo experiment, we used the default design parameters recommended by the designs that are tailored to the "non-informative" case where limited prior knowledge is available on the toxicity profile of the investigational drug. This is appropriate for evaluating and comparing the general performance of the designs

across a variety of toxicity profiles, and for first-in-human drug trials. For the "me-too" or same-family drugs with a better known toxicity profile, the design parameters should be calibrated based on the available prior information to fit the trial under consideration. For example, if the prior information suggests that the investigational drug is relatively safe, we can choose the design parameters that encourage more aggressive dose escalation to find the MTD quickly.

The designs reviewed here focus on single-agent trials and require that before enrolling the next cohort of new patients, patients who were enrolled into the trial have completed their DLT assessment. This requirement is troublesome when toxicity is late-onset or the accrual is fast. Extension of these novel designs, have been developed to address the late-onset toxicity, e.g., the TITE-CRM [10] and data-augmentation CRM [12], and to handle drug combination trials [60, 13, 61, 21, 24]. Recently, Clertant and O'Quigley [46] propose a flexible semiparametric dose finding methods that reduces to the CRM under some added parametric conditions, and is equivalent to the mTPI or BOIN design under some relaxation of the underlying structure. The semiparametric dose finding method shows competitive performance. Comprehensively investigating the existing phase I designs under such a unified framework is of interest and warrants further research.

In chapter 3, we proposed a flexible Bayesian optimal design for phase II trials (BOP2) with simple and complex endpoints under a unified framework, and in chapter 4 we extended it to the BOP2-C design which simultaneously monitors the first cycle tumor response rate and the long-term durable response rate (cure rate). Our BOP2 (BOP2-C) design can explicitly control the type I and II error rates, thereby bridging the gap between Bayesian designs and frequentist designs. In addition, unlike many existing Bayesian designs which use the posterior probability

to make go/no-go decision, the BOP2 design allows the posterior probability cutoff to vary with the interim sample size, which improves the power of the design and reduces the probability of incorrectly terminating the trial early when the treatment is actually promising. Another important feature of the BOP2 design is that the stopping boundaries can be enumerated prior to the onset of trial, making it particularly easy to implement in practice.

The BOP2 design requires the response of each patient can be observed in a very short period after the patient is enrolled into the trial. However, in practice, sometime we need a long time to evaluate the clinical response of patients. If this happens, the patient accrual may be suspended until the results of previous patients are fully observed. In order to shorten the trial time under the circumstances of such delayed responses, we can extend the BOP2 design using the similar method proposed by Cai et al. [62]. Also, extending the BOP2 design from the categorical endpoints to time-to-event endpoints warrants further investigation.

In this dissertation, we studied phase I and phase II clinical trial designs separately. Numerous phase I/II designs have also been proposed to simultaneously consider toxicity and efficacy by combining dose-finding methods and interim decision making together. Thall and Russell (1998) proposed a phase I/II design to characterize the patient outcome with a trinary ordinal variable which account for both toxicity and efficacy [63]. Thall and Cook (2004) introduced a Bayesian phase I/II design based on evaluating the trade-offs between toxicity and efficacy [64]. Yin et al. (2006) proposed a Bayesian dosing-find method using the odds ratios between toxicity and efficacy for phase I/II clinical trials [65]. Yuan and Yin (2009, 2011) developed a phase I/II design to jointly model toxicity and efficacy as time-to-event outcomes, and a Bayesian phase I/II design for drug-combination trials [66, 67]. Liu

et al. recently proposed the first Bayesian phase I/II trial design for immunotherapy, which simultaneously models the immune response, toxicity, and efficacy [68]. The phase I/II trials designs accounting for immunotherapeutic endpoints warrants future work.

Bibliography

[1] Barry E Storer. Design and analysis of phase i clinical trials. *Biometrics*, pages 925–937, 1989.

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

[3] Jeffrey M Skolnik, Jeffrey S Barrett, Bhuvana Jayaraman, Dimple Patel, and Peter C Adamson. Shortening the timeline of pediatric phase i trials: the rolling six design. *Journal of Clinical Oncology*, 26(2):190–195, 2008.

[4] Stephen D Durham, Nancy Flournoy, and William F Rosenberger. A random walk rule for phase i clinical trials. *Biometrics*, 53:745–760, 1997.

[5] Anastasia Ivanova, Aliakbar Montazer-Haghighi, Sri Gopal Mohanty, and Stephen D Durham. Improved up-and-down designs for phase i trials. *Statistics in medicine*, 22(1):69–82, 2003.

[6] Mario Stylianou and Dean A Follmann. The accelerated biased coin up-and-down design in phase i trials. *Journal of biopharmaceutical statistics*, 14(1):249–260, 2004.

[7] John O'Quigley, Margaret Pepe, and Lloyd Fisher. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, 46(1):33–48, 1990.

[8] 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.

[9] B Neuenschwander, M Branson, and T Gsponer. Critical aspects of the bayesian approach to phase i cancer trials. *Statistics in Medicine*, 27(13):2420–2439, 2008.

[10] Ying Kuen Cheung and Rick Chappell. Sequential designs for phase i clinical trials with late-onset toxicities. *Biometrics*, 56(4):1177–1182, 2000.

[11] Guosheng Yin and Ying Yuan. Bayesian model averaging continual reassessment method in phase i clinical trials. *Journal of the American Statistical Association*, 104(487):954–968, 2009.

[12] Suyu Liu, Guosheng Yin, and Ying Yuan. Bayesian data augmentation dose finding with continual reassessment method and delayed toxicity. *Annals of Applied Statistics*, 4:2138–2156, 2013.

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

[14] 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.

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

[16] Fangrong Yan, Sumithra J Mandrekar, and Ying Yuan. Keyboard: a novel bayesian toxicity probability interval design for phase I clinical trials. *Clinical Cancer Research*, 23:3994–4003, 2017.

[17] Yuan Ji, Ping Liu, Yisheng Li, and B Nebiyou Bekele. A modified toxicity probability interval method for dose-finding trials. *Clinical Trials*, 7(6):235–244, 2010.

[18] Wentian Guo, Sue-Jane Wang, Shengjie Yang, Henry Lynn, and Yuan Ji. A bayesian interval dose-finding design addressing ockham's razor: mtpi-2. *Contemporary Clinical Trials*, 58:23–33, 2017.

[19] Suyu Liu and Ying Yuan. Bayesian optimal interval designs for phase i clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(3):507–523, 2015.

[20] Ying Yuan, Kenneth R Hess, Susan G Hilsenbeck, and Mark R Gilbert. Bayesian optimal interval design: a simple and well-performing design for phase I oncology trials. *Clinical Cancer Research*, 22(17):4291–4301, 2016.

[21] Ruitao Lin and Guosheng Yin. Bayesian optimal interval design for dose finding in drug-combination trials. *Statistical methods in medical research*, 26:2155–2167, 2017.

[22] Ruitao Lin and Guosheng Yin. Stein: A simple toxicity and efficacy interval design for seamless phase i/ii clinical trials. *Statistics in medicine*, 36:4106–4120, 2017.

[23] Kentaro Takeda, Masataka Taguri, and Satoshi Morita. Bayesian optimal interval design for dose finding based on both efficacy and toxicity outcomes. *Pharmaceutical Statistics*, page in press, 2018.

[24] Haitao Pan, Ruitao Lin, and Ying Yuan. Statistical properties of the keyboard design with extension to drug-combination trials. *http://arxiv.org/abs/1712.06718*, 2018.

[25] Rongji Mu, Ying Yuan, Jin Xu, Sumithra J. Mandrekar, and Jun Yin Yin. gboin: A unified model-assisted phase i trial design accounting for toxicity grades, binary or continuous endpoint. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, DOI: 10.1111/rssc.12263, 2018.

[26] Peter F Thall and Richard M Simon. Recent developments in the design of phase ii clinical trials. In *Recent Advances in Clinical Trial Design and Analysis*, pages 49–71. Springer, 1995.

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

[28] Thomas R Fleming. One-sample multiple testing procedure for phase ii clinical trials. *Biometrics*, pages 143–151, 1982.

[29] Lisa Garnsey Ensign, Edmund A Gehan, Douglas S Kamen, and Peter F Thall. An optimal three-stage design for phase ii clinical trials. *Statistics in medicine*, 13(17):1727–1736, 1994.

[30] T Timothy Chen. Optimal three-stage designs for phase ii cancer clinical trials. *Statistics in medicine*, 16(23):2701–2711, 1997.

[31] Stephanie J Green and Steve Dahlberg. Planned versus attained design in phase ii clinical trials. *Statistics in medicine*, 11(7):853–862, 1992.

[32] Edward L Korn and Richard Simon. Data monitoring committees and problems of lower-than-expected accrual or event rates. *Controlled clinical trials*, 17(6):526–535, 1996.

[33] Peter F Thall and Richard Simon. Practical bayesian guidelines for phase iib clinical trials. *Biometrics*, pages 337–349, 1994.

[34] Daniel F Heitjan. Bayesian interim analysis of phase ii cancer clinical trials. *Statistics in medicine*, 16(16):1791–1802, 1997.

[35] Say-Beng Tan and David Machin. Bayesian two-stage designs for phase ii clinical trials. *Statistics in medicine*, 21(14):1991–2012, 2002.

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

[37] Ira Mellman, George Coukos, and Glenn Dranoff. Cancer immunotherapy comes of age. *Nature*, 480(7378):480, 2011.

[38] Suzanne L Topalian, George J Weiner, and Drew M Pardoll. Cancer immunotherapy comes of age. *Journal of Clinical Oncology*, 29(36):4828, 2011.

[39] Jennifer Couzin-Frankel. Cancer immunotherapy, 2013.

[40] James C Yang and Richard Childs. Immunotherapy for renal cell cancer. *Journal of Clinical Oncology*, 24(35):5576–5583, 2006.

[41] Suzanne L Topalian, Mario Sznol, David F McDermott, Harriet M Kluger, Richard D Carvajal, William H Sharfman, Julie R Brahmer, Donald P Lawrence, Michael B Atkins, John D Powderly, et al. Survival, durable tumor remission, and long-term safety in patients with advanced melanoma receiving nivolumab. *Journal of clinical oncology*, 32(10):1020, 2014.

[42] Edward B Garon, Naiyer A Rizvi, Rina Hui, Natasha Leighl, Ani S Balmanoukian, Joseph Paul Eder, Amita Patnaik, Charu Aggarwal, Matthew Gubens, Leora Horn, et al. Pembrolizumab for the treatment of non–small-cell lung cancer. *New England Journal of Medicine*, 372(21):2018–2028, 2015.

[43] Helen Gogas, John Ioannovich, Urania Dafni, Catherine Stavropoulou-Giokas, Konstantina Frangia, Dimosthenis Tsoutsos, Petros Panagiotou, Aristidis Poly-

zos, Othonas Papadopoulos, Alexandros Stratigos, et al. Prognostic significance of autoimmunity during treatment of melanoma with interferon. *New England Journal of Medicine*, 354(7):709–718, 2006.

[44] Bethany Jablonski Horton, Nolan A. Wages, and Mark R. Conaway. Performance of toxicity probability interval based designs in contrast to the continual reassessment method. *Statistics in Medicine*, 36(2):291–300, 2017.

[45] Revathi Ananthakrishnan, Stephanie Green, Mark Chang, Gheorghe Doros, Joseph Massaro, and Michael LaValley. Systematic comparison of the statistical operating characteristics of various phase i oncology designs. *Contemporary Clinical Trials Communications*, 5:34–48, 2017.

[46] Matthieu Clertant and John O'Quigley. Semiparametric dose finding methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79:1487–1508, 2017.

[47] Richard E Barlow, David J Bartholomew, JM Bremner, and H Daniel Brunk. *Statistical inference under order restrictions: The theory and application of isotonic regression*. Wiley New York, 1972.

[48] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[49] Heng Zhou, J Jack Lee, and Ying Yuan. Bop2: Bayesian optimal design for phase ii clinical trials with simple and complex endpoints. *Statistics in medicine*, 36(21):3302–3314, 2017.

[50] Bruce D. Cheson, Beate Pfistner, Malik E. Juweid, Randy D. Gascoyne, Lena Specht, Sandra J. Horning, Bertrand Coiffier, Richard I. Fisher, Anton Hagen-

beek, Emanuele Zucca, Steven T. Rosen, Sigrid Stroobants, T. Andrew Lister, Richard T. Hoppe, Martin Dreyling, Kensei Tobinai, Julie M. Vose, Joseph M. Connors, Massimo Federico, and Volker Diehl. Revised response criteria for malignant lymphoma. *Journal of Clinical Oncology*, 25(5):579–586, 2007. PMID: 17242396.

[51] Kathleen N Moore, Michael W Sill, Meaghan E Tenney, Christopher J Darus, David Griffin, Theresa L Werner, Peter G Rose, and Robert Behrens. A phase ii trial of trebananib (amg 386; ind# 111071), a selective angiopoietin 1/2 neutralizing peptibody, in patients with persistent/recurrent carcinoma of the endometrium: An nrg/gynecologic oncology group trial. *Gynecologic oncology*, 138(3):513–518, 2015.

[52] EA1 Eisenhauer, Patrick Therasse, Jan Bogaerts, LH Schwartz, D Sargent, Robert Ford, J Dancey, S Arbuck, S Gwyther, M Mooney, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

[53] Stefano Sacchi, Raffaella Marcheselli, Alessia Bari, Gabriele Buda, Anna Lia Molinari, Luca Baldini, Daniele Vallisa, Marina Cesaretti, Pellegrino Musto, Sonia Ronconi, et al. Safety and efficacy of lenalidomide in combination with rituximab in recurrent indolent non-follicular lymphoma: final results of a phase ii study conducted by the fondazione italiana linfomi. *Haematologica*, pages haematol–2015, 2016.

[54] Peter F Thall, Richard M Simon, and Elihu H Estey. Bayesian sequential monitoring designs for outcomes. *Statistics in medicine*, 14:357–379, 1995.

[55] Peter F Thall and Hsi-Guang Sung. Some extensions and applications of a bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics*

in medicine, 17(14):1563–1580, 1998.

[56] Peter F Thall, Elihu H Estey, and Hsi-Guang Sung. A new statistical method for dose-finding based on efficacy and toxicity in early phase clinical trials. *Investigational new drugs*, 17(2):155–167, 1999.

[57] Ying Yuan, Hoang Q Nguyen, and Peter F Thall. *Bayesian Designs for Phase I–II Clinical Trials*. 2016.

[58] Joseph Berkson and Robert P Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952.

[59] Wally R Gilks, NG Best, and KKC Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics*, pages 455–472, 1995.

[60] 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, 2009.

[61] Liangcai Zhang and Ying Yuan. A practical bayesian design to identify the maximum tolerated dose contour for drug combination trials. *Statistics in Medicine*, 35(27):4924–4936, 2016.

[62] Chunyan Cai, Suyu Liu, and Ying Yuan. A bayesian design for phase ii clinical trials with delayed responses based on multiple imputation. *Statistics in medicine*, 33(23):4017–4028, 2014.

[63] 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.

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

[65] 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.

[66] 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.

[67] Y Yuan and G Yin. Bayesian phase i/ii drug-combination trial design in oncology. *Annals of Applied Statistics*, 5:924–942, 2011.

[68] Suyu Liu, Beibei Guo, and Ying Yuan. A bayesian phase i/ii trial design for immunotherapy. *Journal of the American Statistical Association*, (just-accepted), 2017.

VITA

Heng Zhou was born in Nanjing, Jiangsu, China. After graduating from the High School Affiliated to Nanjing Normal University, he entered the Kuang Yaming Honors School, Nanjing University, Nanjing, China. In June of 2012, he got the B.S. degree in Mathematics concentrated in Statistics. After that, he entered the George Washington University, D.C., USA, and got the M.S degree in Statistics in May of 2014. In August of 2014, he entered The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, for the Ph.D. degree in Biostatistics.