12-2022

Bayesian Adaptive Clinical Trial Design

Mengyi Lu

BAYESIAN ADAPTIVE CLINICAL TRIAL DESIGN

by

Mengyi Lu, M.S.

APPROVED:

_____

Ying Yuan, Ph.D.
Advisory Professor

_____

Yisheng Li, Ph.D.

_____

Suyu Liu, Ph.D.

_____

Ruitao Lin, Ph.D.

_____

Melinda Yates, Ph.D.

_____

APPROVED:

_____

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

BAYESIAN ADAPTIVE CLINICAL TRIAL DESIGN

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

Mengyi Lu, M.S.

Houston, Texas

December, 2022

ACKNOWLEDGEMENT

First of all, I would like to express my most profound appreciation to my advisor, Dr. Ying Yuan, for his expertise, guidance, and invaluable advice. He has supported me at every step, helping me and mentoring me. I could not have made it this far without his generous help and supervision. Additionally, this endeavor would not have been possible without the generous help and support of Dr. Suyu Liu and Dr. Ruitao Lin, both in my research and in my life. I also appreciate the other members of my academic advisory committee: Dr. Yisheng Li and Dr. Melinda S. Yates. Thank you for providing invaluable advice, ideas, and feedback. And thank you for taking the time to help support my research.

I would also like to thank the dean of the GSBS, Dr. William Mattox, for always supporting us and for directing the milestone of our research study. Many thanks to the previous and current QS program directors, Dr. Prahlad Ram, Dr. Liang Li, and Dr. Traver Hart, for always being supportive and assembling this warm QS family. I would also like to thank Amy Carter for always being helpful and supportive. She encouraged me to join the QS student council, which inspired me and helped me step out of my "comfort zone." I would also like to acknowledge my friends and classmates for their encouragement and assistance, as well as the talks, the laughs, and the time we shared.

More importantly, I would like to extend a heartfelt “thank you” to my parents: Thank you for encouraging me to start this unique and remarkable journey. I am deeply grateful for my mother’s unconditional support, encouragement, and love. She is always and forever my most robust supporter. She is always by my side, listen to my negative emotions, always offering positive energy to cheer me up. Without her, I would not have had the courage to embark on this journey in the first place. I also want to extend

exceptional thanks to Dr. Feng Zhang for his many insightful and invaluable suggestions during my Ph.D. journey and in my life. Dr. Zhang has helped me, supported me, and encouraged me at every critical turning point. I would also like to thank my cat Alex for all the emotional support and entertainment he provided during this particular work-from-home time.

ABSTRACT

BAYESIAN ADAPTIVE CLINICAL TRIAL DESIGN

Mengyi Lu, M.S.

Advisory Professors: Ying Yuan, Ph.D.

The landscape of drug development in oncology has changed from conventional chemotherapies to molecular targeted therapies and immunotherapies, which provide innovative therapeutic modalities for treating cancers. These novel therapeutic agents work through mechanisms that fundamentally differ from standard chemotherapeutic agents, making the conventional trial design paradigm inefficient and dysfunctional. Specifically, the focus of dose-finding trials has shifted from finding the maximum tolerated dose (MTD) to the optimal biological dose (OBD), defined as the dose that optimizes the risk–benefit tradeoff. How to accurately identify the OBD and its dosing schedule is of great importance to maximize efficacy and safety of targeted therapies and immunotherapies. The US Food and Drug Administration (FDA) Oncology Center of Excellence recently launched Project Optimus to accelerate this paradigm shift. In addition, once the OBD and recommended phase 2 dose (PR2D) are determined, how to effectively monitor short-term and long-term efficacy in phase II trials, in particular basket trials, is critical for the development of targeted therapies and immunotherapies.

In this dissertation, we propose Bayesian adaptive clinical trial designs to address these challenges. Specifically, we propose (a) a novel Bayesian dose-finding design to find the OBD of drug combination based on risk-benefit tradeoff, (b) a Bayesian adaptive design that simultaneously optimizes dose and schedule based on efficacy, toxicity and PK data, and (c) a phase II basket trial design that uses Bayesian hierarchical model to bor-

row information across treatment arms for efficient termination of ineffective treatment arms based on short-term and long-term endpoints. We conduct extensive simulation studies to evaluate the operating characteristics of the proposed designs. Results show that the proposed designs outperform existing approaches and provide robust and efficient tools to accelerate the development of targeted therapies and immunotherapies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Clinical research consists of research studies or medical research performed on humans as one step in the pharmaceutical development process. A clinical trial (or interventional study) is one type of clinical study in which participants receive specific treatment based on clinical protocols or research plans generated by investigators [1]. Clinical trials aim to evaluate the safety and effectiveness of a new investigational treatment and compare it to current standard treatments. Specifically, in oncology trials, clinical trials aim to determine whether the tumor responds to the investigational treatment (e.g., tumor shrinkage) and evaluate short-term and long-term survival.

The Food and Drug Administration (FDA) defines several discrete phases in a typical series of clinical trials [2]. After drug discovery and animal experiments, the drug development process moves on to clinical research. Phase I clinical trials aim to evaluate the safety of the investigational agent, which can be identified using the maximum tolerated dose (MTD). MTD is defined as the dose-limiting toxicity (DLT) probability of a dose being closest to the target DLT, which refers to the highest acceptable probability of toxicity that is closest to the target toxicity rate. Phase I trials also aim to determine the appropriate dosage and understand how the treatment progresses inside the body.

Phase II clinical trials aim to test the effectiveness of the therapeutic agents, identify side effects, and make "go/no-go" decisions on whether to pursue further investigation in phase III clinical studies. Phase III clinical trials aim to confirm the investigational agent's effectiveness, continue the monitoring of adverse reactions, and compare clinical results to current relevant standard treatments. If the outcome of the phase III clinical

trial is desirable, the FDA will approve the new treatment, and a large-scale phase IV clinical trial will be conducted.

Our research focuses on early-phase (i.e., phase I and phase II) clinical trials.

1.1. Review of previous methods

Extensive statistical methods have been proposed for early-phase clinical trial designs. There are three types of phase I clinical trial designs, based on their respective statistical foundations and implementation: algorithm-based designs, model-based designs, and model-assisted designs.

1.1.1. Algorithm-based designs

Algorithm-based designs are widely used due to their transparency and simplicity. The well-known 3+3 design is one example of this design type. Some extensions of the 3+3 design include the rolling-6 design (Skolnik et al., 2008) and the accelerated titration design (Simon et al., 1997) The 3+3 design was first introduced in the 1940s (Dixon and Mood, 1946), followed by a pharmacological guideline that further described the method in the 1990s . Now, 3+3 is still the most commonly used clinical trial design. From May 31, 2019, to January 1, 2000, the 3+3 design was used in most of the investigator-initiated phase I trials for solid tumors conducted by the Cancer Therapy Evaluation Program (CTEP) sponsored by the National Cancer Institute (Chihara et al., 2022). One primary reason for its ubiquity is that the 3+3 design is simple to understand and implement, and the dose escalation/de-escalation rules are predetermined. However, the 3+3 design is inflexible because it restricts the cohort size to either 3 or 6. The 3+3 design also has poor operating characteristics, shows poor accuracy in identifying the MTD, and often treats patients with lower dose levels due to excessive dose-escalation steps (Le Tourneau et al., 2009).

1.1.2. Model-based designs

The model-based approach is a type of adaptive clinical trial design that was proposed to improve the accuracy of algorithm-based designs. The model-based approach uses statistical methods (e.g., the logistic model) to describe the dose–toxicity curve. One example model-based approach is the continual reassessment method (CRM), which is a Bayesian design that incorporates current dose outcomes as well as all information from enrolled patients. Instead of a prespecified dose-escalation rule, the model-based design specifies a statistical model. If the prespecified model is inappropriate, the CRM design will assign patients to a dose level with relatively higher toxicity.

To overcome this limitation, Babb and Rogatko (Babb et al., 1998; Rogatko et al., 2005) proposed an escalation with overdose control (EWOC) design that employs an alternative Bayesian approach. The EWOC adds assessments for each patient when the dose-limiting toxicity exceeds the MTD. Extensive studies have shown that the CRM design outperforms the 3+3 design by demonstrating higher accuracy in identifying the MTD and more flexibility in enrolling patients. However, the CRM design requires complicated statistical modeling and complex computational implementation that clinicians find difficult to understand, which limits the usefulness of the CRM design in practice.

Some modifications and extensions of the CRM have been proposed to simplify the model. Neuenschwander et al. (2015) proposed a two-parameter Bayesian logistic regression method (BLRM) to utilize escalation. Cheung and Chappell (2000) proposed a time-to-event CRM (TITE-CRM) that considers the time-to-event endpoints for each patient. Yin and Yuan (2009) proposed a Bayesian model that averages the CRM (BMA-CRM) in multiple parallel CRM models with various prespecified toxicity rates. Liu et al. (2013) developed a Bayesian data-augmentation CRM (DA-CRM) to sample model pa-

rameters and missing data.

1.1.3. Model-assisted designs

Recently, the model-assisted approach has been developed by taking advantage of the algorithm-based and model-based methods. Similar to the algorithm-based design, the model-assisted design specifies rules to guide dose escalation/de-escalation before initiating the trial. Similar to the model-based method, the model-assisted approach uses statistical modeling (e.g., the binomial model) to derive efficient real-time decision-making rules. For example, the modified toxicity probability interval design (mTPI) proposed by Ji et al. (2010) uses hierarchical beta-binomial models and Bayesian statistical models. The mTPI design models the probability of toxicity only at the current dose level, and it groups the toxicity rates into three categories (i.e., overdosing, underdosing, and proper dosing) based on the equivalence interval (Kurzrock et al., 2021). Compared to the model-based design, the model-assisted design maintains the same accuracy in identifying the MTD, but it is simpler to understand and implement.

Another representative model-assisted design is the Bayesian optimal interval (BOIN) design proposed by Liu and Yuan (2015), which is a flexible, finite-sample based approach that chooses upper and lower interval boundaries. The BOIN design compares the observed DLT of the current dose to two fixed, predetermined dose escalation/de-escalation boundaries. The BOIN design is more flexible and has superior operating characteristics while offering the same transparency and simplicity as the 3+3 design. The BOIN design is much more accurate in identifying the MTD, and it provides effective controls to assign patients to under- or over-doses.

The BOIN design is more flexible than the 3+3 design in three ways. First, it can target any prespecified DLT rate. Second, the BOIN design has no cohort size

requirements. Third, escalation/de-escalation decision making can occur at any time during the trial (Yuan et al., 2016). Some modifications and extensions of the BOIN design include the generalized BOIN design (gBOIN) proposed by Mu et al. (2019), which incorporates the current toxicity scoring system under a unified framework. Zhou et al. (2019). proposed a utilized BOIN design (U-BOIN) that uses a multinomial-Dirichlet model that evaluates toxicity and efficacy simultaneously and uses the utility function to describe the dose–efficacy tradeoff.

1.1.4. Phase II trial designs

The primary objective of phase II clinical trials is to evaluate the efficacy of the investigational treatment and decide whether the investigational treatment is suitable for large-scale phase III studies. Therefore, the fundamental idea of phase II clinical trial designs is to allow early termination for futility if the treatment cannot attain the prespecified minimum efficacy.

A wide variety of phase II clinical trial designs have been proposed. The most well-known phase II frequentist design is Simon's optimal two-stage design (Simon, 1989) and its extensions (Chen, 1997; Ensign et al., 1994; Hanfelt et al., 1999; Jung et al., 2001; Lin and Shih, 2004; Shuster, 2002). Under the Bayesian framework, some phase II clinical trial designs include those of Tan and Machin (2002), Thall and Simon (1994), Thall et al. (1995), Lee and Liu (2008), Johnson and Cook (2009), Wathen et al. (2008), and Zhou et al. (2017). Compared to phase I clinical trials, phase II clinical trials are more complicated because oncology trials require several efficacy endpoints. Examples include binary outcomes (e.g., tumor shrinkage) and continuous outcomes (e.g., overall survival).

Phase II clinical trial designs, such as our proposed design, commonly utilize

futility monitoring rules to make timely and efficient decisions. These futility monitoring rules usually focus on a single outcome (e.g., the response rate), although a single outcome may not adequately determine the efficacy of the experimental treatment. In actual practice, efficacy response requires more time to observe; therefore, the designers of phase II clinical trials must also consider treatments with continuous endpoints.

1.2. Background and overview

The oncology field dramatically evolved with the introduction of precision medicine. Cancer treatment changed from conventional chemotherapies to novel molecular targeted therapies and immunotherapies. The main objective of early-phase clinical trials has shifted from finding the MTD to finding the optimal biological dose (OBD), defined as the dose that optimizes the risk–benefit tradeoff. This shift is driven by one characteristic of targeted therapy and immunotherapy agents: Their efficacy may not increase with the dose. Thus, the MTD may not be the dose that delivers the optimal therapeutic benefit.

To accelerate this paradigm shift, the US Food and Drug Administration (FDA) Oncology Center of Excellence recently launched Project Optimus. The FDA's release of their Guidance on Benefit-Risk Assessment for New Drug and Biological Products further confirms the stance of the regulatory authority on the importance of optimizing both the dose and administration schedule for novel oncology drugs. Therefore, this optimization is critical for developing novel, adaptive, early-phase clinical trial designs in molecular targeted therapies and immunotherapies.

Sometimes, a single-agent therapy might not provide adequate therapeutic effects for cancer treatment. With the advent of novel molecular targeted therapies and immunotherapies, drug combination therapy is becoming the basis for overcoming cancer resistance and improving treatment efficacy. In addition, the commonly used sequen-

tially outcome-adaptive dose-finding designs assume that patient outcomes are observable shortly after treatment. However, non-cytotoxic therapies often have late-onset outcomes in terms of both toxicity and efficacy. Numerous studies have been proposed, but most designs require complicated statistical modeling, computation, and simulations to demonstrate performance and operating characteristics. Hence, it is critical to propose a robust and easy-to-implement Bayesian adaptive early-phase design for drug combination trials and late-onset outcomes. However, finding the OBD combination (OBDC) in drug-combination trials is more challenging because of the increased dimensionality of the dose-range finding.

In Chapter 2, we extend the single-agent Bayesian optimal interval phase I/II (BOIN12) design to efficiently identify the OBDC in a drug-combination trial. Following the decision-making rules of the BOIN12 design, we assign patients to the most appropriate dose combination by continuously updating the posterior distributions of toxicity and efficacy. When the toxicity or efficacy outcomes are late-onset, we propose a time-to-event version of the design that utilizes patients' follow-up data for decision making. Extensive simulation studies indicate that the proposed extensions of the BOIN12 are more straightforward to implement than the current phase I/II drug-combination designs. The proposed designs also have outstanding operating characteristics for determining the OBDC, based on various trial configurations.

Given a particular dose, the schedule of administration has a profound impact on the drug's toxicity and efficacy profiles. The dosing schedule changes the pharmacokinetics (PK) of the drug. This is one fundamental reason why it also affects the drug's toxicity and efficacy profile. PK describes how the drug is absorbed, distributed, metabolized, and eliminated after administration (Danhof et al., 2005). PK is inherent to dose-schedule optimization because PK dictates how the dose and its administration

schedule affects toxicity and efficacy. Thus, it is also critical to integrate PK data for efficient dose-schedule optimization.

Motivated by this, in Chapter 3, we propose a Bayesian PK-integrated dose-schedule finding (PKIDS) design to identify the optimal dose-schedule regime by integrating PK, toxicity, and efficacy data. Based on the causal pathway through which dose and schedule affect PK—which, in turn, affects efficacy and toxicity—we model the three endpoints jointly by first specifying a Bayesian hierarchical model for the marginal distribution of the longitudinal dose-concentration process. Conditional on the drug concentration in plasma, we model toxicity and efficacy jointly as a function of the concentration. We quantify the risk–benefit of regimes using utility while continuously updating estimates of PK, toxicity, and efficacy based on interim data. Then, we make adaptive decisions to assign new patients to appropriate dose-schedule regimes via adaptive randomization. The simulation study shows that the PKIDS design has desirable operating characteristics.

A basket trial evaluates treatment effects simultaneously in patients with different histologic cancer types and a common biomarker signature. Though basket trials are more resource-intensive, due to its increased efficiency relative to the traditional approach, a series of trials investigating single histologies at a time, basket trial design is rapidly evolving. It is necessary to develop innovative basket trial designs to monitor drug efficacy and stop cohorts that fail to show evidence of activity as early as possible.

To meet this challenge, in chapter 4, we propose a Bayesian hierarchical monitoring design for basket trials by incorporating both short-term endpoints and long-term endpoints. Conditional on latent subgroup indicator, we use the Bayesian hierarchical model to borrow information across different cancer types, increasing efficiency in detecting a

meaningful treatment effect. Extensive simulation studies illustrate that our proposed design has favorable operating characteristics compared with current trial designs. Our proposed Bayesian hierarchical monitoring model yields higher power to detect treatment effects. Simultaneously, it can reduce the probability of early termination when the duration of response is substantially prolonged, but no improvements in response rate.

The design and execution of a clinical trial is a complex process. Differences in various components of the process (e.g., enrollment, eligibility criteria, clinical procedures) could result in discrepancies between the findings of related studies. Potentially exaggerated findings and findings that are contradicted in subsequent studies are not unusual in clinical research. These uncertainties are of particular concern when highly cited clinical studies are involved. Given the widespread impact of highly cited studies on clinical research and practice, these studies are often regarded as models or standards in related research. Thus, a careful statistical analysis of such studies is important because such an analysis helps us understand the process of clinical science and helps clinical researchers appropriately evaluate and interpret experimental findings.

In Chapter 5, we analyze 49 original, highly cited clinical studies that were subsequently contradicted or found to have overestimated the effects of experimental interventions. This analysis is challenging because these studies are highly heterogeneous, and the data retrieved from the corresponding publications are typically limited to summary statistics, without patient-level information. We overcome these difficulties by basing our analysis on test statistics within a Bayesian framework. We identify one source of the contradictory results: the p values strongly overstated the experimental evidence. For highly cited studies, when the p value was .05, there was a 74.4% chance of confirming the null hypothesis. The use of a p value of .05 as the criterion for significance has caused many spurious positive findings that were contradicted by subsequent large-scale studies.

CHAPTER 2

Comb-BOIN12: A Bayesian Phase I/II Trial Design to Find the Optimal Biological Dose for Drug Combination Trials

2.1. Introduction

In early-phase clinical trials, phase I clinical trials aim to determine the maximum tolerated dose (MTD) of the investigational agent. The purpose of phase II clinical trials is to identify the efficacy of the investigational agent. Such traditional objectives and procedures are designed for conventional chemotherapies, such as cytotoxic agents that directly target tumor cells. Promising novel targeted therapy and immunotherapy (e.g., molecular targeted agents, biological agents) focus on the direct pathways of the immune system. These therapies consider the mode of action and continue separately to identify the safety and efficacy of the novel agents inappropriate for novel cancer treatment. Instead, in the era of targeted therapy and immunotherapy, early-phase clinical trials aim to identify the optimal biological dose (OBD), which is the dose that produces the optimal therapeutic effect among all investigational dose levels. Therefore, researchers usually conduct a single study that simultaneously monitors the toxicity and efficacy of phase I/II clinical trials.

Numerous phase I/II trial designs have been proposed to determine the OBD. Zang and Lee investigated a robust two-stage I/II trial design by incorporating toxicity and efficacy outcomes simultaneously (Zang and Lee, 2017). Liu et al. developed a Bayesian trial design by considering the immune response, toxicity, and efficacy outcomes based co-incidentally on unique features of immunotherapy (Liu et al., 2018). Zhou et al. proposed

a utility-based seamless Bayesian phase I/II trial design to determine the OBD by modeling toxicity and efficacy outcomes jointly (Zhou et al., 2019). Lin and Ji investigated a Joint i3+3 design that incorporates toxicity and efficacy outcomes to define the OBD (Lin and Ji, 2020). Lin et al. developed a Bayesian optimal interval phase I/II (BOIN12) design to identify the OBD using a quasi-beta-binomial method (Lin et al., 2020). However, a single agent is insufficient for actual cancer treatment. As a result, combination therapy, which combines two or more therapeutic agents, is becoming foundational for overcoming cancer resistance and improving treatment efficacy(Mokhtari et al., 2017).

The motivation of our design is a combination therapy that utilizes pembrolizumab to treat multiple types of advanced solid tumors. Pembrolizumab is an immunotherapeutic that targets and blocks the programmed cell death receptor (PD-1) immune checkpoint with functional antitumor activity (Robert et al., 2015). In 2014, the FDA approved pembrolizumab for the treatment of advanced melanoma patients who have a BRAF mutation (Ascierto et al., 2012). The drug was later approved to treat several solid tumors that indicate microsatellite instability (MSI-H) and mismatch repair deficiency (Syn et al., 2017). As of 2020, the FDA has also approved pembrolizumab for the first-line treatment of unresectable or metastatic microsatellite instability-high (MSI-H) or mismatch repair deficient (dMMR) colorectal cancer (Food and Administration, 2020).

Although single-agent pembrolizumab showed continued benefit for patients, a dual-type combination of pembrolizumab provides better outcomes for patients who encounter disease relapse as resistance develops (Robert et al., 2015). For example, patients with advanced endometrial cancer or renal cell carcinoma demonstrated promising antitumor activity under safety dosages when treated with the targeted agent lenvatinib in combination with pembrolizumab (Taylor et al., 2020). However, as of December 2020, most phase I/II combination therapy trials involving pembrolizumab still

use conventional clinical paradigms (e.g., the 3+3 design) to identify the MTD before proceeding to the cohort expansion stage (Gangadhar et al., 2015; Mitchell et al., 2018; Kawazoe et al., 2020; Hamid et al., 2017; Smith et al., 2017; Barzi et al., 2022; Pollack et al., 2019; Tawbi et al., 2018; Mato et al., 2018; Aggarwal et al., 2022; Powderly et al., 2020; Johnson et al., 2018; Hirai et al., 2021).

Numerous designs for drug combination trials have been proposed. Yuan and Yin proposed a seamless phase I/II design for drug combination trials using a copula-type regression across the two-dimensional dose–toxicity space in the phase I stage (Yuan and Yin, 2011). Wages and Conaway developed a Bayesian adaptive phase I/II trial design by incorporating two binary responses, toxicity, and efficacy, to assign patients the most optimal dose combination (Wages and Conaway, 2014). Cai et al. investigated a novel Bayesian phase I/II trial design to sufficiently explore untried doses in combination trials (Cai et al., 2014). Guo and Li proposed a Bayesian dose-finding algorithm to find the optimal dose combination without parametric model assumptions (Guo and Li, 2015). Jimenez et al. proposed a seamless two-stage phase I/II trial design with a late-onset efficacy endpoint (Jiménez et al., 2020).

These phase I/II dose-finding methods have greatly improved the development of novel cancer treatments, and they have enhanced the accuracy of OBD determination. However, the designs described above are both conceptually and computationally complex. Most designs require complex statistical model-fitting to determine the OBD, which is both difficult to implement in practice and difficult for clinicians to understand. To overcome these barriers and limitations of the current phase I/II drug combination trial designs, we propose a combination Bayesian optimal interval phase I/II trial design (Comb-BOIN12) for early-phase dose-finding in targeted therapy and immunotherapy. We use a utility function to optimize the risk-benefit trade-offs and to ensure that our

proposed design is simple, accurate, and easy for clinicians to implement in practice.

As an extension of the single-agent BOIN12 design, under the Comb-BOIN12 design, patients are adaptively assigned the most appropriate dose combination by continuously updating the posterior distributions of toxicity and efficacy. Moreover, we find one practical problem with the motivation example: Toxicity and effectiveness have a long response time after treatment begins. With the expected accrual rate, new patients were enrolled in the treatment even though the toxicity and efficacy outcomes of previous patients were still pending. As a result, patients were assigned to unacceptable dose levels.

To address this issue, we also propose a time-to-event version of our design to allow real-time decision-making in cases of late-onset toxicity and efficacy outcomes. Simulation studies indicate that under various configurations and trial settings, our proposed Comb-BOIN12 design is simple to implement, and it exhibits far better operating characteristics for determining the optimal biological dose combination (OBDC) than current phase I/II drug-combination designs.

This chapter is organized as follows: 2.2 describes the BOIN12 design and presents our proposed Comb-BOIN12 designs. 2.3 discusses the results of our extensive simulation studies to evaluate the operating characteristics of our proposed method. 2.4 provides an overview and conclusion.

2.2. Method

Consider the investigation of a single-agent phase I/II trial to determine the OBD with J dose levels, where $J = 1, \ldots, J$. Assume Y_T and Y_E are binary outcomes that evaluate toxicity and efficacy, respectively. When toxicity occurs, $Y_T = 1$; otherwise, $Y_T = 0$. Similarly, $Y_E = 1$ indicates efficacy, while $Y_E = 0$ indicates a lack of efficacy. The liter-

ature includes many proposals to quantify the risk-benefit trade-offs (Thall and Russell, 1998; Gooley et al., 1994; Braun, 2002; Thall and Cook, 2004; Yin et al., 2006; Jin et al., 2014; Guo and Li, 2015; Liu and Johnson, 2016).

Most conventional early-phase trial designs use complicated statistical models to jointly account for toxicity and efficacy, which are difficult to compute. One innovation of our method is that we use utility to quantify risk-benefit trade-offs. Following the BOIN12 approach, we model utility directly, and we use a pseudo-likelihood approach to obtain the posterior utility for decision-making. Given any patient in the trial, all possible outcomes of Y are summarized as $Y = \{(Y_T = 0, Y_E = 1), (Y_T = 0, Y_E = 0), (Y_T = 1, Y_E = 1), (Y_T = 1, Y_E = 0)\} := \{(0,1), (0,0), (1,1), (1,0)\}$. Let p_{ab} present the probability of outcome $(Y_T = a, Y_E = b)$.

Utility $U(Y_T, Y_E)$ should be elicited from clinicians to reflect the risk-benefit trade-off that underlies their medical decisions. The most desirable outcome is $(Y_T = 0, Y_E = 1)$, which is assigned the value of $\rho_{01} = 100$. The least desirable outcome is $(Y_T = 1, Y_E = 0)$, which is assigned a score of $\rho_{10} = 0$. The other two outcomes are suggested by clinicians, and they should fall between the least desirable and most desirable scores. For example, when $\rho_{11} < \rho_{00}$, this utility specification prioritizes efficacy above toxicity because it is desirable to tolerate more toxicity in exchange for increased efficacy. Another advantage of utility is that it is highly scalable and flexible. This utility approach could easily extend beyond bivariate binary endpoints.

2.2.1. Comb-BOIN12 design

Consider a phase I/II combination trial to identify the OBDC from a set of dose combinations $J \times K$. Assume J doses of biological agent A, where $J = 1, dots, J$; and K doses of biological agent B, where $K = 1, \ldots, K$ are involved. Follow the same utility

function strategy to quantify the toxicity-efficacy trade-off, the mean utility is given by

$$u(j,k) = \sum_{a=0}^{1}\sum_{b=0}^{1} u_{ab}p_{ab}(j,k), \tag{2.1}$$

By using quasi-likelihood to model $u(j,k)$ directly, the standardized utility is summarized as

$$u^*(j,k) = u(j,k)/100, \tag{2.2}$$

where $u^*(j,k) \in [0,1]$ is a weighted average of $p_{01}(j,k), p_{00}(j,k), p_{11}(j,k), p_{10}(j,k)$

Under the Bayesian framework, the standard utility follows a beta distribution, and the quasi-binomial likelihood of the observed data iss

$$L(D(j,k)|u^*(j,k)) \propto (u^*(j,k))^{x(j,k)}(1-u^*(j,k))^{n_{jk}-x(j,k)}, \tag{2.3}$$

while the posterior distribution was

$$u^*(j,k)|D(j,k) \sim \text{Beta}(\alpha + x(j,k), \beta + n_{jk} - x(j,k)), \tag{2.4}$$

where $D(j,k) = (n_{jk}, y_{01}(j,k), y_{00}(j,k), y_{11}(j,k), y_{10}(j,k))$, $x(j,k)$ was the number of "events" observed from n_{jk} patients treated at dose (j,k).

Two criteria for evaluating the various dose combinations are used to safeguard patients from overly toxic or futile doses. Incoming patients are treated only under admissible doses, while unacceptable doses are eliminated. Let C_T and C_E indicate the cutoff probabilities, dose combinations are considered unacceptable when they satisfy the

following criteria:

$$\Pr(p_T(j,k) > \phi_T | D(j,k)) > C_T \qquad \text{(Safety)},$$
$$\Pr(p_E(j,k) < \phi_E | D(j,k)) > C_E \qquad \text{(Efficacy)}$$

2.2.2. Dose-finding Algorithm for Comb-BOIN12 design

The rules for dose escalation or de-escalation are easy to follow and understand. Assume that $\hat{p}_T(j,k)$ represents the toxicity rate observed with dose combination (j,k). Let λ_e and λ_d denote the boundaries for escalation and de-escalation adopted from the BOIN design. Let u_b represent the utility benchmark to evaluate $u(j,k)$. Last, assume that N^* presents the sample cutoff, where we recommend $N^* = 6$ for more desirable trial results.

The dose-finding rule is described as follows (see Figure 2.1 for a detailed flowchart):

Stage I (i.e., run-in period)

1. The first cohort of patients is treated with the lowest dose combination $(1,1)$.
2. If no toxicity or efficacy outcome is observed with the current dose combination (j,k), the next cohort of patients is treated with dose combination $(j+1,k+1)$.
3. If $j = k = K$, the dose is increased to $(j+1,K)$. If $j = k = J$, the dose is increased to $(J,k+1)$.
4. Stage I is complete when either toxicity or efficacy is observed.
5. Stage II begins.

Stage II (i.e., dose-finding period)

1. At the current dose combination (j, k), the next cohort of patients is treated based on one of the following three evaluations:

   (a) If $\hat{p}_T(j,k) \geq \lambda_d$, de-escalate one dose level to $(j-1,k), (j,k-1)$

   (b) If $\hat{p}_T(j,k) > \lambda_e$ and $n_{jk} \geq N^*$, the current dose is maintained: (j,k), or $(j-1,k), (j,k-1)$, whichever has the largest value of $\Pr(u_{j'k'} > u_b | D_{j'k'})$

   (c) Otherwise, the next dose combination is chosen from $(j-1,k), (j,k-1), (j,k), (j+1,k), (j,$ whichever exhibits the largest value of $\Pr(u_{j'k'} > u_b | D_{j'k'})$

2. Step I is repeated until the maximum sample size is reached. The final OBD selection is based on the following two procedures:

   (a) MTD is identified as the dose level that shows an isotonically estimated toxicity probability closest to the upper toxicity limit.

   (b) The final OBDC is determined by the dose level that exhibits the highest estimated utility among the doses that do not exceed the MTD.

Another innovation of our proposed design is that we incorporate a model-based approach at Stage II step 2(b), to increase the accuracy of the final OBDC selection. With the same dose-finding rule of Comb-BOIN12, we also propose a Comb-BOIN12$_{\text{MODEL}}$ that uses a model-based approach to model efficacy.

2.2.3. Late-onset Outcomes

Most oncology trial designs assume that toxicity and efficacy outcomes can be observed rapidly after treatment begins. However, one practical problem for targeted therapy and immunotherapy is that toxicity and efficacy are always late-onset factors. Three common late-onset developments occur during novel cancer treatment: (1) a toxi-

city outcome is observed, but efficacy is pending; (2) efficacy is observed, but the toxicity outcome is pending; or (3) both the toxicity and efficacy outcomes are pending. Pending outcomes challenge real-time decision-making, and newly enrolled patients could be assigned inappropriate dosages. Therefore, we also provide a time-to-event version of the Comb-BOIN12 design (TITE-Comb-BOIN12) by adding an approximate likelihood method based on the follow-up time of patients who exhibit late-onset effects at the interim analysis to impute the unobserved missing outcomes.

We assume the observed data (Y_T, Y_E) during the interim time, where $Y_T = 1$ if a patient exhibits dose-limiting toxicity (DLT) at the interim time; otherwise, $Y_T = 0$. Similarly, $Y_E = 1$ if a patient exhibits experimental efficacy at the interim time; otherwise, $Y_E = 0$.

Let $Y_i q, q \in T, E$ denote the toxicity or efficacy outcomes for the ith patient, where $Y_{iq} = 1$ indicates toxicity or efficacy; otherwise $Y_{iq} = 0$. Let ψ_{iq} indicate whether toxicity or efficacy is pending for $Y_{iq}(\psi_{iq} = 0)$ or is observed for $Y_{iq}(\psi_{iq} = 1)$ during the interim time. Patients are divided into four types based on the value of $\psi_q : (\psi_T, \psi_E) = ((1,1),(0,1),(1,0),(0,0))$. The quasi-number of "events" is determined by (a) the observed utility of patients who have both toxicity and efficacy ascertained (x^O) and (b) the observed utility of patients for whom toxicity, efficacy, or both outcomes were pending (x^P).

$$
\begin{aligned}
x = x^O + X^P =& \frac{1}{100} \sum_{a=0}^{1} \sum_{b=0}^{1} \{u_{ab} \sum_{i=1}^{N} I(Y_{iT} = a) I(Y_{iE} = b) \psi_{iT} \psi_{iE} \\
&+ u_{ab} \sum_{i=1}^{N} I(Y_{iT} = a) \Pr(y_{iE} = b | \psi_{iE} = 0) \psi_{iT} (1 - \psi_{iE}) \\
&+ u_{ab} \sum_{i=1}^{N} \Pr(Y_{iT} = a | \psi_{iT} = 0) I(Y_{iE} = b) (1 - \psi_{iT}) \psi_{iE} \\
&+ u_{ab} \sum_{i=1}^{N} \Pr(Y_{iT} = a | \psi_{iT} = 0) \Pr(Y_{iE} = b | \psi_{iE} = 0) (1 - \psi_{iT})(1 - \psi_{iE}) \}
\end{aligned}
\tag{2.5}
$$

When patients have at least one outcome pending, the following assumption is required. For instance, when $Y_{iq} = 1$, the time-to-event outcome t_q is a uniform random variable over $(0, A_q)$, where A_q is the length of the assessment window for Y_q. Then

$$
\begin{aligned}
\Pr(Y_{iq} = 1 | \psi_{iq} = 0) &= \frac{\Pr(\psi_{iq} = 0 | Y_{iq} = 1) \Pr(Y_{iq} = 1)}{\Pr(\psi_{iq} = 1 | Y_{iq} = 0) \Pr(Y_{iq} = 0) + \Pr(\psi_{iq} = 0 | Y_{iq} = 1) \Pr(Y_{iq} = 1)} \\
&= \frac{\Pr(\psi_{iq} = 0 | Y_{iq} = 1) \Pr(Y_{iq} = 1)}{\Pr(\psi_{iq} = 1) + \Pr(\psi_{iq} = 0 | Y_{iq} = 1) \Pr(Y_{iq} = 1)} \\
&= \frac{\pi_q (1 - t_q / A_q)}{1 - \pi_q t_q / A_q} = \frac{\pi_q (1 - \omega_{iq})}{1 - \pi_q \omega_{iq}}
\end{aligned}
\tag{2.6}
$$

Where $\omega_{iq} = \Pr(X_{iq} \leq t | Y_{iq} = 1)$ is the adjusting weight for toxicity ($q = T$) or efficacy ($q = E$) outcomes that remain unobserved. Under the uniform assumption, $\omega_{iq} = t / A_{iq}$.

Given the observed data for N patients, $D_q = \{Y_{iq}, \omega_{iq}, \psi_{iq}, i = 1, \cdots, N\}$. Based on the approximation $1 - \omega_q \pi_q \approx (1 - \pi_q)^{\omega_q}$, the marginal likelihood function for π_q is

given by

$$
\begin{aligned}
L(D_q|\pi_q) &= \prod_{i=0}^{N} [\pi_{iq}^{Y_{iq}}(1-\pi_{iq})^{1-Y_{iq}}]^{\psi_{iq}}(1-\omega_{iq}\pi_{iq})^{(1-\psi_{iq})} \\
&\approx \prod_{i=0}^{N} [\pi_{iq}^{Y_{iq}}(1-\pi_{iq})^{1-Y_{iq}}]^{\psi_{iq}}(1-\pi_{iq})^{\omega_{iq}(1-\psi_{iq}} = \pi_q^{\tilde{v}_q^{(1)}}(1-\pi_{iq})^{\tilde{v}_q^{(0)}+t_q/A_q},
\end{aligned}
\tag{2.7}
$$

During the interim time, $\tilde{v}_q^{(1)} = \sum_i^N \psi_{iq}(1-Y_{iq})$ indicates the number of patients who experienced DLT, $\tilde{v}_q^{(0)} = \sum_i^N \psi_{iq}(1-Y_{iq})$ denotes the number of patients who completed the DLT assessment window but have no experienced DLT; and t_q/A_q represents the standard total follow up time (STFT) of patients who have pending toxicity outcomes. The probability of toxicity and efficacy is estimated as $\hat{\pi}_q = \tilde{v}_q^{(1)}/ESS$, where $ESS = \tilde{v}_q^{(1)} + \tilde{v}_q^{(0)} + t_q/A_q$. The quasi number of events x is calculated by inserting $\hat{\pi}_q$ into the expression $\Pr(Y_{iq} = 1|\psi_{iq} = 0)$. When x can be determined, the Comb-BOIN12 and Comb-BOIN12$_{\text{MODEL}}$ can be applied directly to obtain the posterior of utility for decision making. The dose-finding algorithm for hte time-to-event version follows the same pattern as the Comb-BOIN12 design.

2.3. Simulation Studies

We conduct extensive simulations to investigate the operating characteristics of the proposed Comb-BOIN12 and Comb-BOIN12$_{\text{MODEL}}$ designs under various trial settings. We compare the results from the Comb-BOIN12 and Comb-BOIN12$_{\text{MODEL}}$ designs to the copula-type model, the change-point model, and the Bayesian hierarchical model designs.

The copula-type model design, initially proposed by Yin and Yuan, is a seamless phase I/II dose combination trial. Their design uses a copula-type regression to model toxicity and to select acceptable preliminary doses for phase I. When phase I is complete, this set of acceptable preliminary doses will move seamlessly to phase II, where

patients are allocated to different treatment arms based on their acceptable doses. A novel adaptive randomization procedure is used to differentiate the treatments based on their efficacy levels (Yuan and Yin, 2011).

Cai et al. propose a change-point model-based design (Cai et al., 2014). The run-in period focuses on the exploration of possible dose combinations and the collection of doses with acceptable toxicity and efficacy for further investigation in stage II. Then, a beta-binomial model is implemented to examine the safety requirements. The trial moves on to stage II when the highest dose combination is attained or the safety requirement is violated. In stage II, a change-point model is used in the dose-toxicity surface, and a logistic regression model with quadratic terms is adapted for the non-monotonic dose-efficacy pattern. Based on the toxicity and efficacy observed in stage I, patients are assigned an optimal dose combination by continuously updating the posterior estimation of toxicity and efficacy.

Yada and Hamada propose a Bayesian hierarchical model design that divides phase I/II trials into two stages. Similar to the copula-type model, the dose-escalation rule in stage I is based on a copula-type model. In stage I, a set of acceptable preliminary doses is selected when the maximum sample size is reached. In stage II, a Bayesian hierarchical model is adapted to examine dose–efficacy and dose–toxicity relationships, and patients are assigned using the Bayesian moving reference adaptive randomization method proposed by Yin and Yuan (Yuan and Yin, 2011; Yada and Hamada, 2018).

We conducted multiple trials with varied configurations to investigate the accuracy and reliability of five trial designs: Comb-BOIN12, Comb-BOIN12$_{\text{MODEL}}$, a copula-type model, a change-point model, and a Bayesian hierarchical model. A total of 19 scenarios with various dose–toxicity and dose–efficacy curve shapes were simulated. The

dose–efficacy curves include a plateau shape, an umbrella shape, and a linear shape. All trials included two biological agents: drug A and drug B.

We used the same scenarios and followed the same trial settings presented by the Bayesian hierarchical model (Scenarios 1–9 in Table 2.1). Then, we extracted representative scenarios from the change–point model (Scenarios 10–15 in Table 2.4) and the copula-type model (Scenarios 16–19) Table 2.5. In all designs, the trial begins with the lowest dose level (A_1, B_1).

The maximum sample size of Scenarios 1–9 is 51, with a cohort size of three. We consider four dose levels for drug A and drug B, respectively. The upper toxicity limit is $\phi_T = 0.35$, and the efficacy lower limit is $\phi_E = 0.2$.

Scenarios 10–15 adopt the same dose level for drug A and drug B, but the maximum sample size is 45 in a cohort size of three. The highest acceptable toxicity upper limit is $\phi_T = 0.3$, and the acceptable efficacy lower limit is $\phi_E = 0.2$.

Scenarios 16–19 utilize three dose levels for drug A and two for drug B. The maximum sample size is 42 in a cohort size of three. The upper toxicity limit is $\phi_T = 0.33$, and the efficacy lower limit is $\phi_E = 0.2$.

The cutoff value for toxicity monitoring is $C_T = 0.85$ for dose $(1, 1)$ in order to apply more stringent safety monitoring at the starting dose. The cutoff value for all other doses is $C_T = 0.95$. The cutoff value for futility monitoring is $C_E = 0.9$. The dose-escalation and de-escalation boundaries λ_e and λ_d are derived using the default specifications $\phi_1 = 0.6\phi_T$ and $\phi_2 = 1.4\phi_T$. To intensify the toxicity–efficacy trade-off, we use the utility value presented in Table 2.2. For Comb-BOIN12$_{\text{MODEL}}$, we employed the five-parameter logistic model with quadratic terms to incorporate complicated dose–efficacy

relationships. If we assume that q_{jk} is the probability of a response from dose combination (j, k), the logistic model is

$$\text{logit}(q_{jk}) = \gamma_0 + \gamma_1 a_j + \gamma_2 b_k + \gamma_3 a_j^2 + \gamma_4 b_k^2$$

We set prior of parameters $\gamma_0 \sim \ Cauchy(0, 10)$, $\gamma_i \sim \ Cauchy(0, 2.5)$, and $i = 1, 2, 3, 4$.

We used these 19 representative scenarios to the compare operating characteristics of the designs given various dose–efficacy and dose–toxicity relationships. In Scenarios 1, 5, and 16–19, efficacy monotonically increased with the dose for both drug A and drug B. In Scenario 5, three dose combinations showed optimal desirability. Alternatively, in Scenario 13, efficacy monotonically decreased with the dose for both drugs. Scenarios 4, 6, 8, 10, and 12 showed a non-monotonic dose–efficacy relationship for both drugs. For example, in Scenario 4, the efficacy plateaued at the highest dose level, and the true efficacy probability for drug A in Scenario 8 follows a decreasing curve. In Scenarios 2 and 11, only drug A followed a monotonic drug-efficacy pattern. However, in Scenario 7, only drug B followed a monotonic dose–efficacy relationship. In regard to the dose–toxicity relationship, all dose levels in Scenario 9 exceed acceptable toxicity levels. In Scenarios 10 and 14, both drugs' dose–toxicity surfaces plateau at the highest dose level. In Scenario 19, all true toxicity probability remains the same for both drugs. In the other scenarios, the dose–toxicity relationship monotonically increases for both drugs.

2.3.1. Simulation results

To evaluate the operating characteristics, we examine the following three performance metrics based on 5,000 simulated trials: the probability of selecting the correct OBDC, the percentage of patients treated with the OBDC, and the percentage of patients treated at overdose levels (i.e., toxicity rate (ϕ_T)). The percentage of correctly

selected OBDCs is used to evaluate accuracy when determining the desired dose combination for the trial design. The percentage of patients treated at the OBDC is used to evaluate patient allocation. Higher percentages and higher numbers illustrate better operating characteristics for identifying the desired dose combination. The percentage of patients treated for overdose is used to assess the safety of the design. A lower percentage indicates better performance of the designed trial.

Table 2.3 summarizes the simulation results. We only present the percentage of selection of OBDC of the Bayesian hierarchical model due to missing details in the simulation settings. Both the Comb-BOIN12 and Comb-BOIN12$_{\text{MODEL}}$ designs exhibit overall operating characteristics that far exceed the other three methods in most cases, especially in terms of patient allocation. For instance, in Scenario 5, both methods have higher accuracy rates and patient allocation values when multiple OBDCs occurred. In Scenario 6, both methods show outstanding accuracy and patient allocation performance compared to the other three methods.

Moreover, our proposed designs are more robust and more stable than the other five designs. Both methods produce the same patient allocations and the same number of patients treated at risk dose levels in all scenarios. However, Comb-BOIN12$_{\text{MODEL}}$ identified a higher percentage of correct OBDCs than Comb-BOIN12. Copula-type and change-point models perform better only when the optimal dose combination is on the edge of the two-dimensional dose combination space. For example, in Scenario 1, the optimal dose combination is (a_4, b_4). Copula-type and change-point models have a higher selection percentage, and the copula-type model has a better patient allocation in the OBDC in Scenario 1. However, these models performed worse in the other scenarios overall.

Our proposed methods produce the most robust and stable performance metrics in all scenarios. The performance of the copula-type model provides strong evidence in favor of this conclusion. In Scenarios 1, 2, and 8, the copula-type model is more likely to identify the correct OBDC, but an incorrect OBDC is selected in Scenario 4. The copula-type model demonstrates the maximum optimal dose selection in Scenario 2; however, both of our proposed methods performed superior patient allocation. The change-point model showed higher accuracy only in Scenarios 1 and 2, and it produced lower patient allocation values in most scenarios. In addition, more patients were treated at overdose levels. In most scenarios, the Bayesian hierarchical model was consistently the lowest performer overall. Under the extreme circumstances of Scenario 9, all methods will abort the trial based on established safety procedures.

2.3.2. Sensitivity analyses

We also conducted sensitivity analysis to investigate the performance of our proposed designs under various utility specifications and different sample sizes.

1. *Different utility specifications*

We considered two cases: Case 1, where $u_{01} = 100, u_{00} = 30, u_{11} = 70, u_{10} = 0$ and Case 2, where $u_{01} = 100, u_{00} = 20, u_{11} = 80, u_{10} = 0$. Figure 2.2 shows the simulation results, there were slight variations in (a) the percentage of correct OBDCs selected, (b) the percentage of patients treated with the OBDC, and (c) the number of patients treated with overdoses. However, the simulation performance remained the same. Our proposed design is capable of adjusting the dose-assignment distribution adaptively under various utility specifications.

2. *Different sample sizes*

Figure 2.3 shows the results of the simulation (only the first four scenarios of the Comb-BOIN12 design are presented in this section). The results show that our proposed methods exhibit relatively similar performance when the same four evaluation metrics are used. We chose a minimum sample size of 42 and a maximum of 120. The results illustrate the superior performance of our proposed method when the sample size increases. Moreover, an increased sample size increases the probability of determining the OBDC and improves patient allocation at the OBDC. The percentage of patients treated at overdose levels increases when the total sample size is no more than 57, then it decreases as the sample size increases. This slight increase occurs when the maximum sample size is relatively small because the dose space is explored more fully as the sample size increases. This increases the probability of treating patients at overdose levels, but it is insufficient to accurately identify excess toxicity at these dosage levels. Thus, our methods exhibit superior operating characteristics when the sample size increases, when multiple OBDCs occur, and under various trial configurations. Our sensitivity analysis shows that our proposed method is reliable for practical use.

2.3.3. Late-onset Outcomes

For the TITE-Comb-BOIN12 and TITE-Comb-BOIN12$_{\text{MODEL}}$ designs, we examine the operating characteristics under simulation studies. The same scenario configurations are used in all 19 scenarios. The accrual rate is two patients per month. We use two toxicity and efficacy assessment windows. In the first assessment window, toxicity and efficacy duration are one month and two months, respectively (hereafter, assessment window 1). The second assessment window for toxicity is two months and three months of efficacy (hereafter, assessment window 2). Likewise, the operating characteristics are also evaluated using four performance metrics. Instead of assessing the percentage of

patients treated with the OBDC, we consider the trial duration in order to evaluate its efficiency. For trial duration, a lower value indicates better performance.

For both the assessment windows described above, we compared the TITE-Comb-BOIN12 and TITE-Comb-BOIN12MODEL designs to our proposed Comb-BOIN12 design. The TITE-Comb BOIN12 and TITE-Comb-BOIN12MODEL designs produce similar results. Figure 2.4 summarizes the simulation results. Both methods showed relatively similar performance in terms of OBDC selection accuracy, patient allocation, and the number of patients treated at overdose levels. However, TITE-Comb BOIN12 significantly outperforms Comb-BOIN12 in terms of trial duration in all scenarios.

A significant difference is indicated when the extended toxicity and efficacy assessment window. For example, in assessment window 1 of Scenario 1, TITE-Comb BOIN12 requires 11.5 fewer months than Comb-BOIN12. However, in assessment window 2 of Scenario 1, TITE-Comb BOIN 12 requires 19.3 fewer months than Comb-BOIN12. Similar results occur when multiple OBDCs are included (Scenario 5). In the extreme case, all dose levels are excessively toxic (Scenario 9). In Scenarios 10–19, we examine the performance of TITE-Comb BOIN12 in various configurations. We obtain the same results when we change the sample size or the toxicity limit. For example, in Scenario 13, TITE-Comb BOIN12 required 12.5 fewer months than Comb-BOIN12 in assessment window 1 and 22.2 fewer months in assessment window 2. We also obtained the same result after changing the doses of the biological agents' drug A and drug B. For example, in Scenario 19, TITE-Comb BOIN12 requires 19.8 fewer months than Comb-BOIN12 in assessment window 2, and 8 fewer months than Comb-BOIN12 in assessment window 1. Compared to Comb-BOIN12, our time-to-event version design has similar overall accuracy and patient allocation performance, and it is more efficient when toxicity and efficacy outcomes have longer assessment windows.

2.4. Discussion

We propose a combination Bayesian optimal interval phase I/II trial design (Comb-BOIN12) for novel cancer treatments in early-phase dose-finding. The essential motivation of our proposed design is to overcome the conceptual and computational difficulties that arise due to the non-monotonic efficacy patterns of targeted therapy and immunotherapy in drug combination trials. We use a quasi-beta-binomial approach to simplify clinical implementation, determine the OBDC, and factor in the risk-benefit trade-offs that inform decision-making. One advantage of using a utility function to reflect the risk-benefit trade-off is its high scalability and flexibility. The utility approach can easily extend beyond bivariate binary endpoints. The only modification required is to expand the dimensions of the utility table to cover the outcome space. For example, for three endpoints with r_1, r_2, and r_3 categories, a utility table with the dimensions $r_1 \times r_2 \times r_3$ will cover the outcome space. We also propose a model-based version for drug combination trials to improve accuracy when determining efficacy.

Numerous simulation studies show that our proposed designs surpass the operating characteristics of other existing methods. Our proposed methods are more accurate when selecting an OBDC, produce better patient allocation, and are less likely to treat patients using excessively toxic dose combinations. Our proposed designs also excel in terms of robustness and reliability. Our sensitivity analysis demonstrates the overall remarkable performance of our proposed methods.

Based on these remarkable results, we recommend Comb-BOIN12 for practical use because it is simple, robust, easy to implement, and easy to understand. In addition, we also propose a time-to-event version of Comb-BOIN12 for use with late-onset toxicity and efficacy outcomes. The time-to-event version allows for real-time decision-making for

newly accrued patients by adopting an approximate likelihood approach. TITE-Comb-BOIN12 is similar to Comb-BOIN12 in terms of OBDC accuracy, patient allocation, and overdose control; however, TITE-Comb-BOIN12 is superior in terms of reducing trial duration.

Table 2.1: True probabilities of toxicity and efficacy at each dose combination of scenario 1-9. Boldface indicates the toxicity probability and efficacy probability of the optimal biological dose combination (OBDC).

Scenario	Drug A	Drug B			
		1	2	3	4
1	1	(.05, .16)	(.10, .20)	(.12, .24)	**(.15, .35)**
	2	(.07, .18)	(.11, .22)	(.14, .28)	(.17, .40)
	3	(.10, .20)	(.12, .26)	(.15, .32)	(.19, .45)
	4	(.12, .23)	(.15, .28)	(.16, .35)	(.23, .60)
2	1	(.08, .05)	(.10, .20)	(.12, .22)	(.14, .24)
	2	(.10, .10)	(.12, .25)	(.14, .28)	(.16, .30)
	3	(.12, .15)	(.15, .30)	(.18, .40)	(.21, .32)
	4	(.15, .20)	(.18, .35)	**(.20, .60)**	(.30, .40)
3	1	(.10, .20)	(.18, .20)	(.20, .40)	(.28, .42)
	2	(.20, .24)	(.22, .25)	**(.24, .55)**	(.50, .58)
	3	(.28, .28)	(.35, .30)	(.48, 0.58)	(.55, .62)
	4	(.32, .32)	(.45, .35)	(.56, .62)	(.60, .65)
4	1	(.08, .10)	(.14, .15)	(.18, .25)	(.35, .55)
	2	(.10, .12)	(.16, .20)	(.21, .30)	(.42, .55)
	3	(.12, .16)	(.18, .26)	**(.24, .55)**	(.45, .55)
	4	(.16, .20)	(.20, .30)	(.33, .55)	(.60, .55)
5	1	(.12, .20)	(.16, .25)	(.18, .30)	**(.20, .50)**
	2	(.14, .25)	(.18, .30)	**(.20, .50)**	(.45, .60)
	3	(.18, .30)	**(.20, .50)**	(.42, .60)	(.55, .66)
	4	(.20, .40)	(.40, .60)	(.52, .65)	(.60, .70)

Continued on next page

Table 2.1 – *Continued from previous page*

Scenario	Drug A	Drug B			
		1	2	3	4
6	1	(.18, .20)	(.20, .25)	**(.24, .50)**	(.30, .40)
	2	(.20, .24)	(.24, .30)	(.33, .40)	(.35, .35)
	3	(.25, .28)	(.30, .32)	(.35, .35)	(.42, .30)
	4	(.28, .30)	(.35, .35)	(.40, .30)	(.54, .25)
7	1	(.07, .15)	(.09, .20)	(.11, .20)	(.13, .30)
	2	(.10, .20)	(.11, .25)	(.14, .30)	(.15, .40)
	3	(.12, .25)	(.13, .30)	(.17, .40)	**(.19, .60)**
	4	(.14, .15)	(.15, .20)	(.20, .25)	(.22, .40)
8	1	(.01, .20)	(.05, .20)	(.12, .15)	(.15, .15)
	2	(.03, .30)	(.08, .25)	(.16, .25)	(.20, .20)
	3	**(.06, .55)**	(.12, .35)	(.20, .30)	(.26, .25)
	4	(.10, .35)	(.15, .25)	(.23, .20)	(.33, .15)
9	1	(.50, .52)	(.56, .62)	(.65, .70)	(.68, .76)
	2	(.55, .55)	(.62, .66)	(.70, .74)	(.72, .79)
	3	(.60, .58)	(.67, .70)	(.75, .78)	(.79, .82)
	4	(.62, .65)	(.72, .74)	(.80, .80)	(.85, .85)

Table 2.2: Utility Table

	Efficacy	**No Efficacy**
No Toxicity	100	10
Toxicity	90	0

Table 2.3: Summary of simulation for scenario 1-9

Designs	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	Percentage of correct selection of the OBDC								
Comb-BOIN12	38.60	46.26	32.96	27.32	44.50	26.36	46.64	31.14	5.40
Comb-BOIN12$_{\text{MODEL}}$	43.32	49.86	31.88	24.98	44.60	23.48	49.88	29.96	6.82
Copula-type model	63.42	51.80	7.36	0.00	31.16	7.12	21.30	68.52	0.20
Change-point model	53.78	41.26	14.94	11.44	28.38	16.36	40.70	28.38	2.32
Bayesian hierarchical model	42.80	39.70	16.20	19.50	34.80	17.60	21.70	36.6	n/a
	Percentage of patients treated at the OBDC								
Comb-BOIN12	22.78	26.38	20.55	19.97	35.37	15.19	25.70	12.92	11.50
Comb-BOIN12$_{\text{MODEL}}$	22.78	26.38	20.55	19.97	35.37	15.19	25.70	12.92	11.50
Copula-type model	50.02	10.18	4.02	2.18	14.86	3.25	4.00	18.27	14.29
Change-point model	22.33	19.03	8.55	10.55	19.20	8.59	19.04	12.29	8.80
	Number of patients treated at overdoses								
Comb-BOIN12	0	0	9.82	7.92	14.01	4.36	0	0	51
Comb-BOIN12$_{\text{MODEL}}$	0	0	9.82	7.92	14.01	4.36	0	0	51
Copula-type model	0	0	9.84	13.36	14.19	6.65	0	0	8.77
Change-point model	0	0	18.00	15.67	24.87	6.47	0	0	8.60

Table 2.4: True probabilities of toxicity and efficacy at each dose combination of scenario 10-19. Boldface indicates the toxicity probability and efficacy probability of the optimal biological dose combination (OBDC).

Scenario	Drug A	Drug B			
		1	2	3	4
10	1	(.02, .08)	(.04, .10)	(.07, .29)	(.12, .42)
	2	(.04, .23)	(.08, .28)	(.13, .42)	**(.18, .60)**
	3	(.09, .14)	(.15, .14)	(.18, .24)	(.25, .43)
	4	(.14, .10)	(.25, .10)	(.25, .18)	(.25, .24)
11	1	(.01, .05)	(.07, .22)	(.12, .10)	(.18, .08)
	2	(.05, .12)	(.12, .29)	(.15, .15)	(.20, .10)
	3	(.10, .19)	(.15, .44)	(.19, .20)	(.23, .18)
	4	(.15, .42)	**(.18, .60)**	(.21, .38)	(.25, .32)
12	1	(.05, .30)	(.10, .40)	**(.18, .60)**	(.25, .37)
	2	(.10, .20)	(.15, .28)	(.23, .37)	(.42, .26)
	3	(.15, .10)	(.23, .14)	(.42, .24)	(.43, .18)
	4	(.23, .05)	(.42, .08)	(.43, .15)	(.44, .10)
13	1	**(.05, .60)**	(.18, .37)	(.26, .30)	(.38, .24)
	2	(.15, .37)	(.26, .26)	(.40, .20)	(.49, .13)
	3	(.25, .24)	(.42, .18)	(.46, .14)	(.51, .10)
	4	(.39, .15)	(.45, .10)	(.50, .08)	(.55, .05)
14	1	(.09, .00)	(.17, .11)	(.22, .23)	(.24, .11)
	2	(.17, .01)	(.22, .14)	(.24, .26)	(.25, .13)
	3	(.22, .01)	(.24, .23)	(.25, .39)	(.25, .22)
	4	(.24, .06)	(.25, .44)	**(.25, .62)**	(.25, .42)

Continued on next page

Table 2.4 – *Continued from previous page*

Scenario	Drug A	Drug B			
		1	2	3	4
15	1	(.00, .06)	(.01, .45)	**(.07, .63)**	(.24, .45)
	2	(.01, .03)	(.07, .32)	(.23, .50)	(.39, .32)
	3	(.06, .00)	(.22, .08)	(.38, .18)	(.44, .08)
	4	(.21, .00)	(.38, .00)	(.44, .01)	(.45, .00)
16	1	(.05, .10)	(.15, .30)	**(.20, .50)**	
	2	(.10, .20)	(.15, .40)	(.45, .60)	
17	1	(.05, .10)	(.15, .30)	(.40, .50)	
	2	(.10, .20)	**(.20, .40)**	(.50, .55)	
18	1	(.05, .10)	(.10, .20)	(0.15, 0.40)	
	2	(.10, .20)	(.15, .30)	**(.20, .50)**	
19	1	(.05, .10)	(.05, .20)	(.05, .40)	
	2	(.05, .20)	(.05, .30)	**(.05, .50)**	

Table 2.5: Summary of simulation for scenario 10-19

Designs	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
					Percentage of correct selection of the OBDC					
Comb-BOIN12	38.64	38.22	40.10	71.94	34.80	40.44	43.76	31.12	47.54	53.98
Comb-BOIN12$_{\text{MODEL}}$	45.18	43.36	40.10	78.48	42.24	44.92	51.28	30.98	56.72	64.72
Copula-type model	29.90	0.62	9.22	29.08	15.26	0.76	33.3	29.9	48.4	56.24
Change-point model	37.6	37.72	35.58	71.06	41.08	52.24	40.96	30.42	62.06	64.22
					Number of patients treated at the OBDC					
Comb-BOIN12	8.47	8.04	8.16	26.68	8.63	7.77	10.79	10.37	13.27	15.65
Comb-BOIN12$_{\text{MODEL}}$	8.47	8.04	8.16	26.68	8.63	7.77	10.79	10.37	13.27	15.65
Copula-type model	9.15	0.07	4.16	9.96	1.95	3.22	6.88	9.15	10.01	11.57
Change-point model	8.08	7.96	7.16	17.23	8.03	10.55	10.46	8.83	15.11	15.50
					Percentage of patients treated at the OBDC					
Comb-BOIN12	18.83	17.87	18.13	59.29	19.17	17.27	25.68	24.68	31.59	37.25
Comb-BOIN12$_{\text{MODEL}}$	18.83	17.87	18.13	59.29	19.17	17.27	25.68	24.68	31.59	37.25
Copula-type model	20.33	0.16	9.24	22.13	4.33	7.16	16.38	21.79	23.83	27.55
Change-point model	17.95	17.68	15.91	38.29	17.84	23.44	24.90	21.02	35.97	36.90
					Number of patients treated at overdoses					
Comb-BOIN12	0	0	10.08	7.34	0	10.75	8.65	14.03	0	0
Comb-BOIN12$_{\text{MODEL}}$	0	0	10.08	7.34	0	10.75	8.65	14.03	0	0
Copula-type model	0	0	10.89	12.61	0	15.25	8.21	13.16	0	0
Change-point model	0	0	9.12	10.19	0	9.87	12.80	18.77	0	0

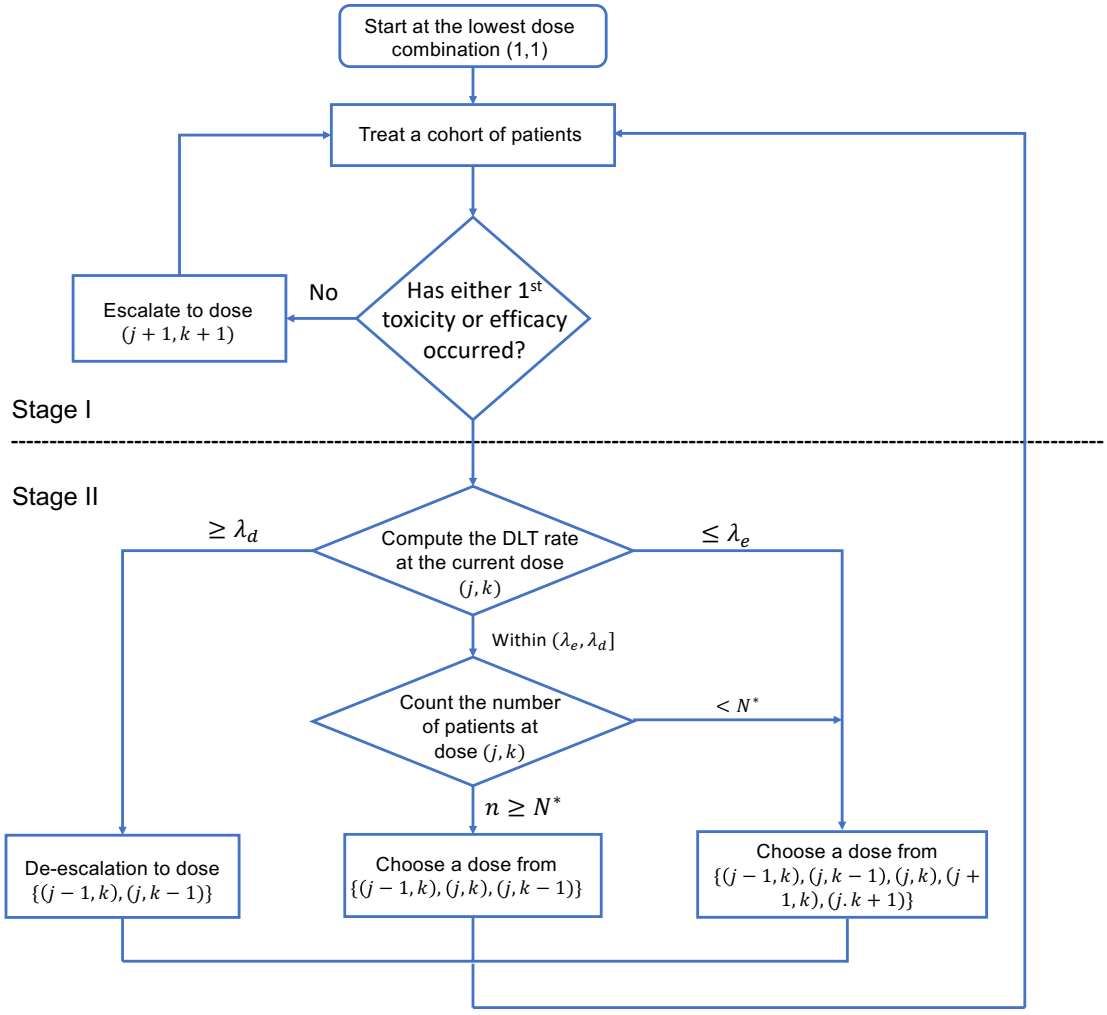

Figure 2.1: Dose finding rule of Comb-BOIN12 design

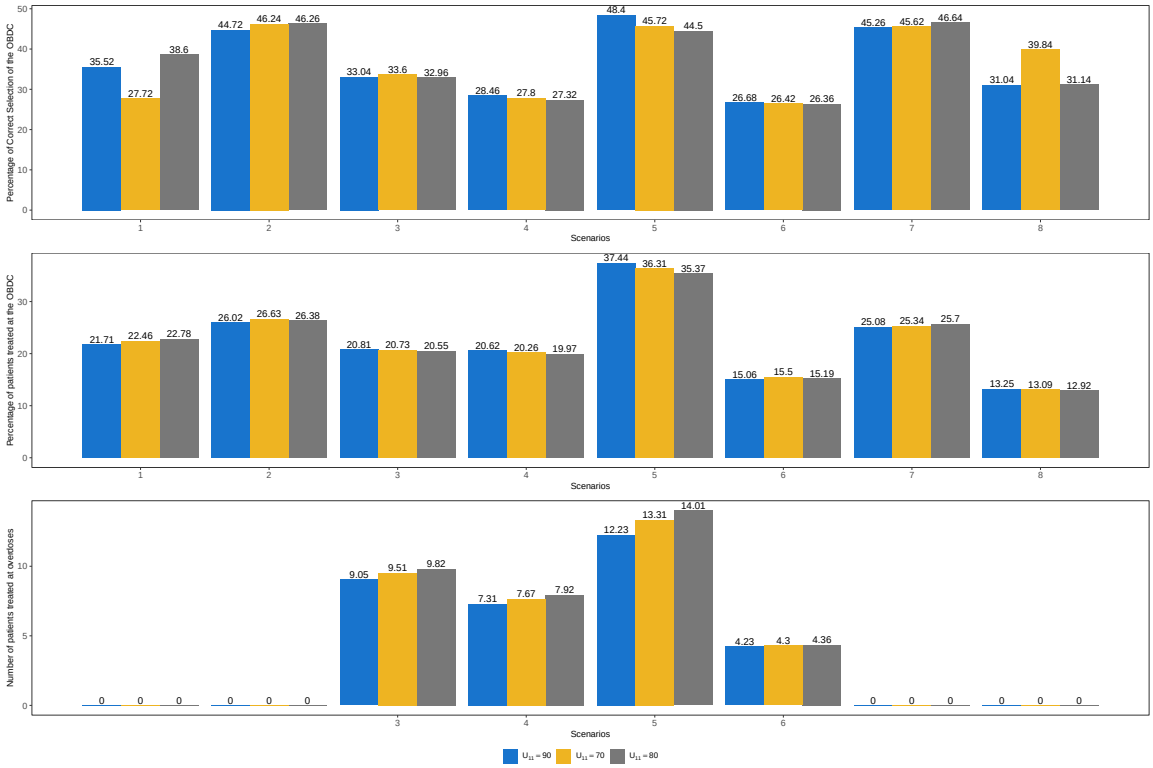

Figure 2.2: Sensitivity Analysis of Comb-BOIN12 (Different utility specifications)

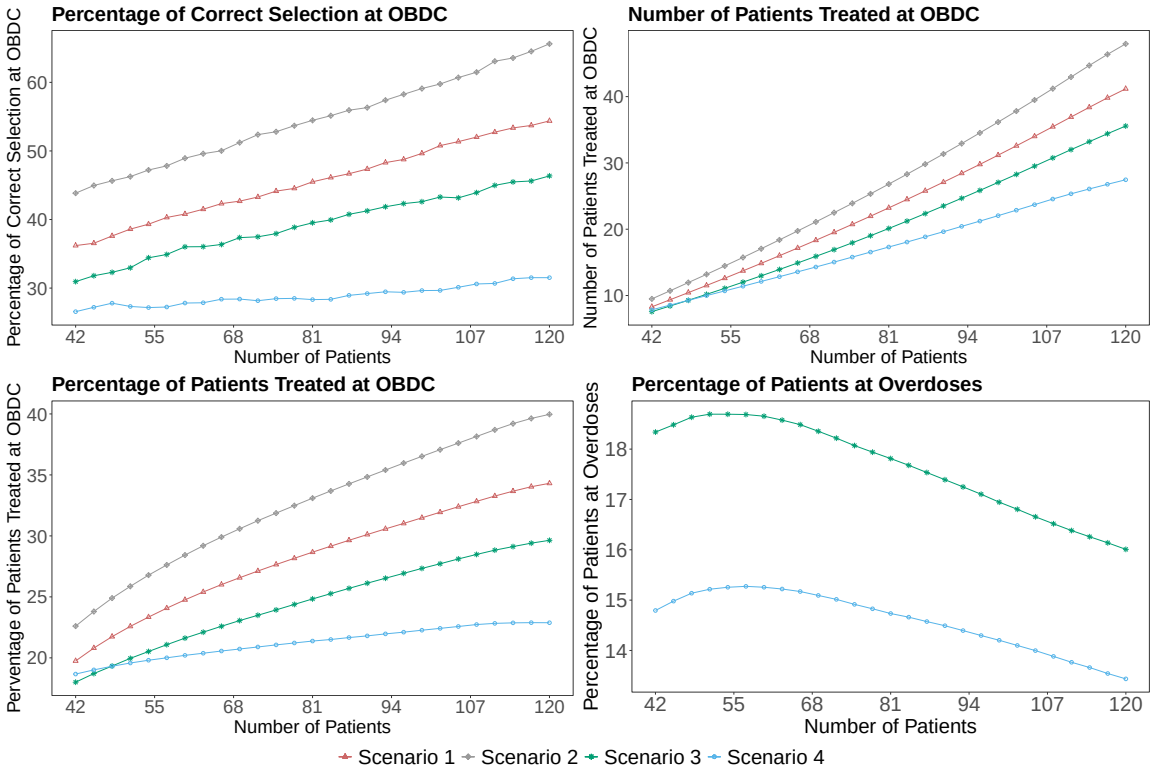

Figure 2.3: Sensitivity Analysis of Comb-BOIN12 (Different sample sizes)

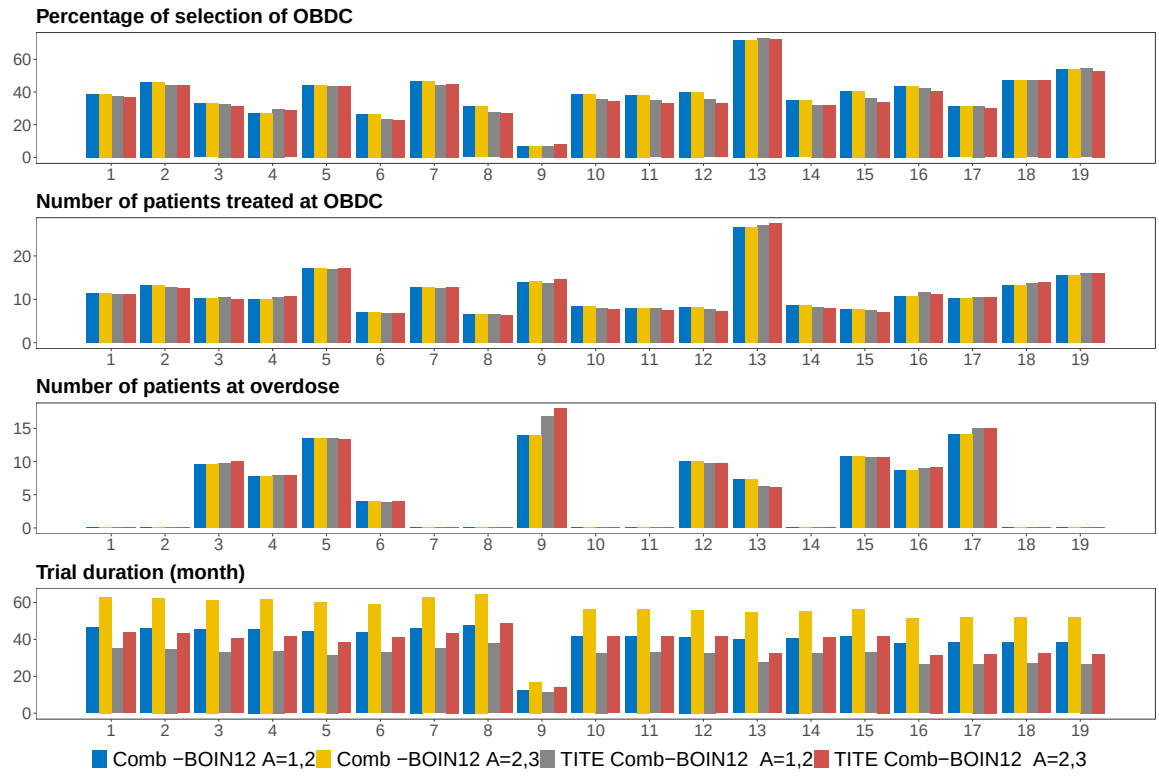

Figure 2.4: Simulation Result of TITE-Comb-BOIN12

CHAPTER 3

A Bayesian Pharmacokinetics Integrated Phase I-II Design to Optimize Dose-schedule Regimes*

3.1. Introduction

In the era of molecularly targeted therapies and immunotherapies, the focus of dose finding trials has been shifting from finding the maximum tolerated dose (MTD) to the optimal biological dose (OBD), defined as a dose that optimizes the risk-benefit tradeoff. The shift is driven by the characteristic of targeted and immunotherapy agents that their efficacy may not increase with the dose, and thus the MTD may not be the dose delivering the optimal therapeutic benefit. The US Food and Drug Administration (FDA)'s Oncology Center of Excellence recently launched Project Optimus to accelerate this paradigm shift. The release of Guidance on Benefit-Risk Assessment for New Drug and Biological Products by the FDA further confirms the stance of the regulatory authority on the importance of optimizing the dose for novel oncology drugs. Many phase I-II designs have been proposed to find OBD by jointly considering toxicity and efficacy, for example, Braun (2002), Thall and Cook (2004), Yuan and Yin (2009), Jin et al. (2014), Liu and Johnson (2016), Takeda et al. (2018), Zhou et al. (2019), Lin et al. (2021), among others. Yuan et al. (2016) and Yan et al. (2018) provide comprehensive reviews on phase I-II design paradigm and methodology.

Given a dose, the schedule of administering the drug has profound impact on its toxicity and efficacy profiles (Blomqvist et al., 1993; Gyergyay et al., 2009; Motzer et al., 2012). For example, gemtuzumab ozogamicin was the first antibody–drug conjugate

*Chapter 3 has been under review in Biostatistics.

approved by the FDA for treating relapsed acute myeloid leukemia in 2000. Gemtuzumab ozogamicin later was withdrawn from the US market in 2010 when a confirmatory trial showed that it was associated with a greater rate of fatal toxicities versus standard-of-care chemotherapy (Petersdorf et al., 2009; Godwin et al., 2017). In 2017, by utilizing a different dosing schedule that improved the safety profile, gemtuzumab ozogamicin was reapproved for relapsed/refractory acute myeloid leukemia after a phase III trial with a fractionated dosing schedule (Jen et al., 2018). As another example, in 2006 the FDA granted accelerated approval of dasatinib administered at 70 mg twice daily for the treatment of chronic myelogenous leukemia with resistance or intolerance to prior therapy. The subsequent confirmative phase III trial found that, compared to 70-mg twice-daily dosing schedule, 100 mg once daily produced similar clinical benefits with significantly lower incidences of key treatment-related adverse events (Shah et al., 2008). This led to the FDA granting full approval in 2009 with a modified label of 100 mg once daily. These examples demonstrate the importance of optimizing both dose and administration schedule to achieve the optimal risk-benefit profile.

A number of dose-schedule finding designs have been proposed. Braun et al. (2007) developed a time-to-event model and design to find the maximum tolerated dose-schedule for cytotoxic agents, based on dose limiting toxicity. Zhang and Braun (2013) extended that method to incorporate adaptive variations to dose-schedule assignments within patients as the trial progresses. Thall et al. (2007) developed a Bayesian phase I-II design to find the optimal dose-schedule (ODS) regime based on a utility that quantifies the toxicity-efficacy tradeoff. Li et al. (2008) introduced a phase I-II dose-schedule finding design that jointly models toxicity and efficacy using Bayesian hierarchical model. Guo et al. (2016) modeled toxicity and efficacy jointly as a trinary endpoint and developed a Bayesian phase I-II design to identify ODS. Cunanan and Koopmeiners (2017)

presented a two-stage, Bayesian phase I-II trial design to optimize the schedule of therapeutic cancer vaccines based on toxicity and immune response.

A fundamental reason that dosing schedule affects the toxicity and efficacy profiles of a drug is that it changes the pharmacokinetics (PK) of the drug. PK is a branch of pharmacology and an indispensable component of drug discovery and development (Danhof et al., 2005). PK describes what the body does to a drug after administration through the mechanisms of absorption, distribution, metabolism, and elimination. PK is innate to dose-schedule optimization, as the effects of the dose and its administration schedule on toxicity and efficacy is executed by PK. Thus, it is of intrinsic interest and importance to integrate PK data for efficient dose-schedule optimization. Most existing dose-schedule finding designs, however, largely ignore PK data. Günhan et al. (2020) employed a PK model to describe the DLT event process, but it considers only toxicity, ignoring efficacy endpoints, and thus is not suitable to find ODS for targeted and immune therapies. Without considering the schedule, a number of dose finding designs have been proposed to utilize PK data to achieve more efficient dose finding, including Piantadosi and Liu (1996), Whitehead et al. (2007), Ursino et al. (2017), Günhan et al. (2021), among other.

In this paper, we propose a Bayesian PK integrated dose-schedule finding (PKIDS) design to identify ODS by integrating PK, toxicity, and efficacy data. Based on the causal pathway, we jointly model these three endpoints by first specifying a Bayesian hierarchical model for the marginal distribution of the longitudinal dose-concentration process. Conditional on the concentration of the drug in plasma, we jointly model toxicity and efficacy as a function of the concentration. We continuously update the estimates of PK, toxicity, and efficacy based on interim data, and we make adaptive decisions to assign new patients to appropriate dose-schedule regimes via adaptive randomization. The

simulation study shows that the PKIDS design has desirable operating characteristics.

Our research is motivated by a trial to establish the optimal dose-schedule regime for a novel agent targeting the yes-associated protein (YAP) in patients with advanced solid tumors. YAP is a downstream target of the Hippo pathway, a key signaling pathway involved in the regulation of organ size (Tumaneng et al., 2012) and playing a role in tumorigenesis (Lee et al., 2010). Five doses (80, 200, 400, 600, and 900 mg) and two administration schedules (administer the dose every 4 days for a total of 7 times, or administer the dose in half every 2 days for a total of 14 times, for a 28-day treatment cycle) will be studied. Toxicity will be scored as a binary endpoint using NCI Common Terminology Criteria for Adverse Events version 5.0. Efficacy will be scored as a binary endpoint (response/no response) using Response Evaluation Criteria in Solid Tumors (RECIST) version 1.1. Collection of samples to assess the plasma concentration of the drug will be collected at pre-dose and 1, 2, 2.5, 3, 4, 6, and 8 hours from the start of infusion.

This paper is organized as follows. 3.2 proposes the joint probability model for PK, toxicity and efficacy data. Section 3.3 describes the utility approach to quantifying the toxicity-efficacy tradeoff, and the dose-schedule finding algorithm based on the utility. Section 3.4 presents simulation studies and the operating characteristics of the proposed PKIDS design. Conclusions and discussions are provided in Section 3.5.

3.2. Methods

Consider a trial aiming to identify the ODS from a set of prespecified $J \times K$ dose-schedule pairs (d_j, s_k), $j = 1, \cdots, J$, $k = 1, \cdots, K$. Let Y_T denote the binary toxicity outcome, with $Y_T = 1$ indicating toxicity (or severe adverse events), and Y_E denote the binary efficacy outcome, with $Y_E = 1$ indicating favorable response. Depending on the

trial, Y_E can be tumor response or biological activity of the drug measured by biomarkers, e.g., pharmacodynamic (PD) endpoint or other surrogate efficacy endpoint. Let Z denote the PK endpoint, representing the plasma concentration of the drug, which is measured longitudinally. The outcome used for dose-schedule finding in our design is a trivariate mixture of longitudinal and scalar endpoints (Z, Y_T, Y_E). In contrast, most existing dose-schedule finding designs are either based on Y_T or (Y_T, Y_E) that are scalar. To the best of our knowledge, this is the first dose-schedule finding design that considers such a longitudinal-scalar trivariate mixture data structure. Adaptive decisions of the PKIDS design (e.g., dose-schedule assignment and selection) are based on the relationship of (Z, Y_T, Y_E) as a function of (d_j, s_k).

Another innovation of our method is the incorporation of the knowledge of the causal path among (Z, Y_T, Y_E) to guide the modeling strategy. As shown in Figure 2.1, the effects of (d_j, s_k) on the therapeutic outcomes (Y_T, Y_E) are mediated by the plasma concentration of the drug in the body (i.e., Z). This motivates us to factorize the joint distribution of (Z, Y_T, Y_E) as

$$[Z, Y_T, Y_E | d_j, s_k] = [Z | d, s][Y_T, Y_E | Z, d_j, s_k] \tag{3.1}$$

In what follows, we first describe the marginal model for $[Z|d_j, s_k]$, followed by the conditional model of (Y_T, Y_E), given Z.

3.2.1. PK Model

Suppose patient i receives $d_{[i]}$ with schedule $s_{[i]} = (m_i, \tau_i)$, under which a fractional dose d_i/m_i is administered every τ_i hours for a total of m_i times, where $d_{[i]} \in d_1, \cdots, d_J$ and $s_{[i]} \in s_1, \cdots, s_K$. Let $Z_{il} \equiv Z(t_{il})$ denote the observed (plasma) concentration of the drug for patient i at the lth time point, t_{il}, $l = 1, \cdots, L$. We assume a one-

compartment model with first-order absorption and first-order elimination, characterized by three PK parameters: absorption rate k_a, elimination rate k_e, and compartment volume V (Davidian and Gallant, 1992; Jones and Rowland-Yeo, 2013). This model assumes that the rate of drug absorption from the gut is proportional to the amount of drug in the gut with proportionality constant k_a, and the rate of elimination from the plasma compartment is proportional to the amount of drug in the plasma compartment with proportionality constant k_e. The compartment volume V is a proportionality constant, relating the total drug in the plasma compartment to the concentration in that compartment. To facilitate model specification, we reparameterize (k_a, k_e, V) as (ϕ, ψ, δ), where $\phi = \log k_a$, $\psi = \log(k_e V)$ and $\delta = \log V$, so that the support of the parameters is the whole real line, where $k_e V$ is often known as the clearance parameter in pharmacology.

Following Wakefield (1996) andMeibohm and Derendorf (1997), we employ a three-level Bayesian hierarchical model to describe the relationship between $Z(t_{il})$ and $(d_{[i]}, s_{[i]})$. Let i index the ith patient specific PK parameters. The first level of the model hierarchy specifies the patient-level relationship between $Z(t_{ij})$ and $(d_{[i]}, s_{[i]})$:

$$\log Z(t_{il}) = \log f(t_{ij}|\phi_i, \psi_i, \delta_i, d_{[i]}, s_{[i]}) + \varepsilon_{il} \tag{3.2}$$

$$\begin{aligned} &f(t_{ij}|\phi_i, \psi_i, \delta_i, d_{[i]}, s_{[i]}) \\ &\quad = \frac{\exp(\phi_i) F \frac{d_{[i]}}{m_i}}{\exp(\delta_i + \phi_i) - \exp(\psi_i)} \left\{ \frac{\exp(-\exp(\psi_i - \delta_i)t)}{1 - \exp(-\exp(\psi_i - \delta_i))\tau_i} \right. \\ &\quad \left. - \frac{\exp(-\exp(\phi_i)t)}{1 - \exp(-\exp(\phi_i))\tau_i} \right\} \end{aligned} \tag{3.3}$$

where ε_{il} is independent and identically distributed, following the normal distribution $N(0, \sigma^2)$, and F is bioavailability, presenting the fraction of the administered dose that reaches the measurement site. F typically is assumed to be known or estimated from external data. We here assume bioavailability F as unity, which is standard practice

when no reasonable value for F is available. The specification of patient-specific PK parameters (i.e., ϕ_i, ψ_i, δ_i) acknowledges the heterogeneity across patients. In practice, there is often little data available in the absorption phase immediately after the drug administration, making it difficult to estimate patient-specific absorption parameter ϕ_i. Therefore, we assume that ϕ_i is the same for all individuals, i.e., $\phi_i = \phi$.

The second-level hierarchy specifies the distribution of patient-specific PK parameters to relate them to population PK parameters. Since ϕ is assumed to be the same across patients, we here only need to consider ψ_i and δ_i, modeled as follows:

$$\begin{pmatrix} \psi_i \\ \delta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right), \tag{3.4}$$

where $\exp(\mu_1)$ and $\exp(\mu_2)$ are the population clearance parameter and compartment volume, respectively.

The third-level hierarchy specifies the prior of the model parameters, including σ^2, ϕ, μ_1, μ_2 and Σ. We adopt a vague prior for σ^2, μ_1, μ_2, and Σ as follows:

$$\sigma^2 \sim IG(0.001, 0.001),\ \mu_1 \sim N(0, 10^3),\ \mu_2 \sim N(0, 10^3) \tag{3.5}$$

$$\Sigma \sim IW\left(\begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}, 2 \right), \tag{3.6}$$

where $IG(a, b)$ denotes inverse gamma distribution with shape parameter a and scale parameter b, and $IW(Q, \nu)$ denotes an inverse Wishart distribution with scale matrix Q and ν degrees of freedom. The specification of prior for ϕ needs more consideration, as in

practice $k_a > k_e$. To reflect this constraint, we assign ϕ a truncated normal distribution,

$$\phi \sim N(\mu_\phi, \sigma_\phi^2) I\left(\phi > \max_i(\phi_i - \delta_i)\right) \tag{3.7}$$

where $I(\cdot)$ is the indicator function, and μ_ϕ and σ_ϕ^2 are hyperparameters. As described previously, data contain limited information on ϕ, and a conventional noninformative prior with a large variance often leads to unstable posterior inference. Therefore, we adopt a weakly informative prior, determined as follows: elicit a plausible range of k_a from subject matter experts, say $[L_{ka}, U_{ka}]$, and choose the values of μ_ϕ and σ_ϕ^2 such that 95% of the prior density fall within $[L_{ka}, U_{ka}]$.

The PK model describes the concentration of the drug over time. A standard summary measure to summarize the cumulative exposure of the drug over time is the area under the curve (AUC). Under the proposed model, the population average AUC at (d_j, s_k) is given by

$$AUC_{jk} = \int_0^T f(t|\phi, \mu_1, \mu_2, d_j, s_k)dt \tag{3.8}$$

where $T = m_i\tau_i$ is the duration of drug administration, and $f(t|\phi, \mu_1, \mu_2, d_j, s_k)$ is the concentration curve evaluated at the population PK parameters, given by

$$\begin{aligned} &f(t|\phi, \mu_1, \mu_2, d_j, s_k) \\ &\quad = \frac{\exp(\phi) F \frac{d_j}{m_k}}{(\exp(\mu_2 + \phi) - \exp(\mu_1))} \left\{ \frac{\exp(-\exp(\mu_1 - \mu_2)t)}{1 - \exp(-\exp(\mu_1 - \mu_2)\tau_k)} \right. \\ &\quad \left. - \frac{\exp(-\exp(\phi)t)}{1 - \exp(-\exp(\phi)\tau_k)} \right\} \end{aligned} \tag{3.9}$$

AUC is not the only choice. When appropriate, other commonly used PK statistics include the maximum concentration (C_{max}) and half-life time $t_{1/2}$, can also be included in the regression model described below to quantify the effects of the plasma concentration

of the drug on clinical endpoints.

3.2.2. Toxicity and Efficacy Model

This section specifies the distribution of $[Y_T, Y_E|Z, d_j, s_k]$. Conditional on Z, let $p_{E,jk}$ and $p_{T,jk}$ denote the probability of efficacy and the probability of toxicity, respectively, for the dose-schedule pair (d_j, s_k), i.e., $p_{E,jk} = \Pr(Y_E = 1|d_j, s_k, Z)$ and $p_{T,jk} = \Pr(Y_T = 1|d_j, s_k, Z)$. We assume Y_E and Y_T marginally follow a logistic model

$$\begin{aligned} \text{logit}(p_{E,jk}) &= \alpha_0 + \alpha_1 AUC_{jk} \\ \text{logit}(p_{T,jk}) &= \beta_0 + \beta_1 AUC_{jk} \end{aligned} \tag{3.10}$$

where α_0, α_1, β_0 and β_1 are regression parameters. In this model, we assume that the effect of (d_j, s_k) on Y_T and Y_E are fully mediated by the cumulative exposure of the drug. That is, (d_j, s_k) together affects the concentration of the drug in the body, which in turn results in clinical responses such as toxicity and efficacy. In contrast, almost all existing dose-schedule finding methods regard d_j and s_k as two independent covariates in their models, without accounting for the close interplay between d_j and s_k and the underlying causal path. When appropriate, $\text{C}_{\max}$ and other PK measures can also be included as covariates in the model.

In the above model, we assume that both toxicity and efficacy increases with the exposure (e.g., AUC). However, as the PK model (3) allows the exposure to plateau with the dose d, the efficacy model above somewhat accommodates the case that efficacy Y_E plateaus with d. Nevertheless, when appropriate, a plateaued logistic model (Cai et al., 2014) such as $\text{logit}(p_{E,jk}/\varrho) = \beta_0 + \beta_1 AUC_{jk}$ can be used, where ϱ represents where the exposure-efficacy curve plateaus.

Given the marginal model, we model the joint distribution of (Y_E, Y_T) using the

Gumbel model as follows:

$$\begin{aligned}\pi_{a,b}(j,k) &= \Pr(Y_{E,jk}=a, Y_{T,jk}=b) \\ &= p_{E,jk}(1-p_{E,jk})^{1-a} p_{T,jk}(1-p_{T,jk})^{1-b} \\ &+ (-1)^{a+b} p_{E,jk}(1-p_{E,jk}) p_{T,jk}(1-p_{T,jk}) \frac{\exp(\gamma)-1}{\exp(\gamma)+1}\end{aligned} \tag{3.11}$$

where $a, b \in (0,1)$, and $\gamma \geq 0$ is the association parameter, introducing the correlation between Y_T and Y_E.

To specify the prior distribution for the parameters that appear in $[Y_T, Y_E|Z, d_j, s_k]$, we take the regularized weakly informative prior approach (Gelman et al., 2008; Guo and Yuan, 2017), such that the resulting prior is vague enough to cover the plausible values of the parameter, but not too vague as to cause stability issues due to sparse data of early phase trials. Under logistic model, a change of 5 on the logit scale moves the probability of the outcome variable from 0.05 to 0.89 or from 0.5 to 0.99, which covers the plausible range of the effect size of a covariate on toxicity and efficacy probabilities. Therefore, we scale the input variables (i.e., AUC_{jk}) to have mean 0 and standard deviation (SD) 0.5, and assign α_1 and β_1 an independent normal prior $N(0, 2.5^2)$, such that a change in any of these covariates from one SD below the mean to one SD above the mean most likely results in a difference of less than 5 (i.e., 2 SD of the prior) on the logit scale.

To set a prior for α_0 and β_0, we elicit from clinicians the upper limit for toxicity rate $\bar{p}_T$ and the lower limit for efficacy rate $\underline{p}_E$, then we assign prior $\alpha_0 \sim N(\hat{\alpha}_0, \hat{\alpha}_0^2)$ and $\beta_0 \sim N(\hat{\beta}_0, \hat{\beta}_0^2)$, with $\hat{\alpha}_0 = \text{logit}^{-1}(\underline{p}_E)$ and $\hat{\beta}_0 = \text{logit}^{-1}(\bar{p}_T)$ respectively, with a coefficient of variation = 1, which are spread out enough to cover the range of $p_{E,jk}$ and $p_{T,jk}$ in practice. We assign γ a uniform prior $\gamma \sim Unif(0,2)$ to cover the practically realistic range of the correlation between Y_E and Y_T.

3.2.3. Likelihood and Posterior

Suppose that n subjects have been treated in the trial, the observed data are $D = \{(Z(t_{ij1}), \cdots, Z(t_{ijL}), Y_{E,i}, Y_{T,i}), i = 1, \cdots, n\}$. Let θ denote the collection of model parameters. The likelihood of D is given by

$$\begin{aligned} L_{jk}(D|\theta) &= \prod_{i=1}^{n} \left\{ L_{ijk}(Y_{ijk}^E, Y_{ijk}^T|\alpha, \beta, \varphi) \times L_{ijk}(Z_{ijk}|\phi, \psi_i, \mu, \sigma, \sigma^2) \right\} \\ &= \prod_{i=1}^{n} \Bigg\{ \{\pi_{00}^{ijk}\}^{(1-Y_{ijk}^E)(1-Y_{ijk}^T)} \{\pi_{10}^{ijk}\}^{Y_{ijk}^E(1-Y_{ijk}^T)} \{\pi_{01}^{ijk}\}^{1-Y_{ijk}^E Y_{ijk}^T} \{\pi_{11}^{ijk}\}^{Y_{ijk}^E Y_{ijk}^T} \\ &\quad \times 2\pi|\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} \frac{(\log Z_{ijkt} - \log f(\psi_i, t))^2}{\sigma^2} - \frac{1}{2}(\psi_i - \mu)^T \Sigma^{-1} (\psi_i - \mu) \right\} \Bigg\} \end{aligned} \tag{3.12}$$

Let $\pi(\theta)$ denote the prior distribution of θ; the posterior distribution of θ is given by $\pi(\theta|D) \propto L(D|\theta)\pi(\theta)$. We sample this posterior distribution using Gibbs' sampler based on the adaptive rejection metropolis sampling (ARMS) algorithm (Gilks et al., 1995).

3.3. Dose-Schedule Finding Algorithm

3.3.1. Desirability of Dose-schedule

For each individual endpoint Y_E and Y_T, the evaluation of the desirability of a dose-schedule regime is straightforward. We prefer a regime that has low toxicity and high efficacy. However, when we consider Y_E and Y_T simultaneously, we need to consider the risk-benefit tradeoffs, as physicians routinely do in almost all medical decisions when selecting a treatment for a patient. A convenient tool to formalize such a process is to use a utility function $U(Y_E, Y_T)$ to map the bivariate outcomes into a single index to measure the desirability of a regime in terms of the risk-benefit tradeoffs. This approach

has been used in previous trial designs, see for example Houede et al. (2010), Thall et al. (2013), Guo and Yuan (2017), Murray et al. (2018), Liu et al. (2018) Lin et al. (2021), among others.

Utility $U(Y_E, Y_T)$ should be elicited from clinicians to reflect the risk-benefit trade-off underlying their medical decisions, which can be done using the following procedure:

a) Assign the least desirable outcome $(Y_E, Y_T) = (0, 1)$ the lowest utility score of $\rho_{01} = 0$, and the most desirable outcome $(Y_E, Y_T) = (1, 0)$ the highest utility score of $\rho_{10} = 100$. This sets up the scale of the score system.

b) Ask the clinician to use these two utilities as a reference to assign each of $(Y_E, Y_T) = (0, 0)$ and $(1, 1)$ a score between 0 and 100 to reflect their clinical desirability. Table 3.1 shows the utility elicited for the solid tumors trial, with $\rho_{00} = 35$ and $\rho_{11} = 60$. As $\rho_{11} > \rho_{00}$, this utility specification gives efficacy higher priority than toxicity, i.e., it is desirable to tolerate more toxicity in exchange of efficacy.

Then, the desirability or (mean) utility of (d_j, s_k) is given by

$$U_{jk} = \sum_{a,b \in (0,1)} \rho_{ab} \pi_{ab}(j, k), \tag{3.13}$$

which can be estimated based on interim data D by

$$\hat{U}_{jk} = \sum_{a,b \in (0,1)} \rho_{ab} E(\pi_{ab}(j, k)|D) \tag{3.14}$$

In our experience, clinicians quickly understand what the utilities mean and provide values for $u_{ab}, a, b \in (0, 1)$, since the values reflect the actual clinical practice. After completing this process, simulation should be performed to evaluate the operating char-

acteristics of the design and reviewed with clinicians. In some cases, the simulation results may motivate modification of some of the numerical utility values to better reflect clinical decisions. One possible criticism for using the utility values is that they require subjective input. However, we are inclined to view this as a strength, rather than a weakness. The process of specifying the utility requires clinicians to carefully consider the potential risks and benefits of the treatment that underlie their clinical decision making in a more formal way and incorporate that into the trial. The specification of the utility allows researchers to perform simulation to formally evaluate the impact of a specific risk-benefit tradeoff criteria on the operating characteristics of the design, rather than making a leap of faith that the implicit risk-benefit tradeoff criteria underlying clinical decisions would lead to desirable operating characteristics. More importantly, based on the simulation results, researchers could further calibrate the design to incorporate additional clinical considerations and better achieve trial objectives. In addition, many studies (Guo and Yuan, 2017; Liu et al., 2018; Zhou et al., 2019; Lin et al., 2021), as well as the sensitivity analysis described later, show that the adaptive dose-finding designs are generally not sensitive to the numerical values of the utility, as long as it reflects a similar trend.

Another advantage of utility is that it is highly scalable and flexible. The utility approach can easily extend beyond bivariate binary endpoints. The only modification is to expand the dimension of the utility table (e.g., Table 1) to cover the outcome space. For example, for three endpoints with r_1, r_2, and r_3 categories, a $r_1 \times r_2 \times r_3$ utility table will cover the outcome space. The utility is flexible in that it contains some commonly used tradeoff criteria as a specifical case. Zhou et al. (2019) and Lin et al. (2021) proved that for the bivariate binary case, when setting $u_{10} + u_{01} = 100$, the utility function approach is equivalent to the tradeoff based on the marginal toxicity and efficacy probability, e.g.,

$p_{E,jk} - wp_{T,jk}$, where w is a weight presenting toxicity and efficacy tradeoff.

To safeguard from treating patients at a toxic or futile regime, we define the admissible dose-schedule set $\mathcal{A}$ as the dose-schedules satisfying the following safety and efficacy criteria:

$$\begin{aligned} &\text{(Safety)} \quad \Pr(p_{T,jk} < \bar{p}_T | D) > p_T^* \\ &\text{(Efficacy)} \quad \Pr(p_{E,jk} > \underline{p}_E | D) > p_E^* \end{aligned} \tag{3.15}$$

where (p_T^*, p_E^*) are the probability cutoffs calibrated by simulation, e.g., $p_T^* = p_E^* = 0.05$. In the trial, only regimes in $\mathcal{A}$ can be used to treat patients.

3.3.2. Dose-Schedule Finding Algorithm

Suppose patients are treated sequentially in cohorts of size g, with the maximum sample size of R cohorts, i.e., $N_{max} = Rg$. Our sequential dose-schedule finding algorithm is defined as follows:

Step 1: Equally randomize the first Kg patients to (d_1, s_1), $(d_1, s_2), \cdots, (d_1, s_K)$, such that one cohort of patients are treated at the lowest dose for each schedule $k = 1, \cdots, K$.

Step 2: Suppose r cohorts of patients have been treated, $r = K, \cdots, R$. Let $d_{j_k^*}$ denote the highest tried dose under schedule s_k, and define the dose-schedule exploration set $\mathcal{H} = \{(d_j, s_k); \Pr(p_{T,j_k^* k} < \bar{p}_T | D) > p_s^* \ \& \ j_k^* < J\}$, where p_S^* is a probability cutoff. $\mathcal{H}$ represents the set of dose-schedules, for which the highest tried dose is sufficiently safe such that the exploration of the next higher dose is warranted. Based on interim data D, determine the admissible dose-schedule set $\mathcal{A}$ and exploration set $\mathcal{H}$. If $\mathcal{A}$ is empty, terminate the trial and declare no ODS; otherwise, assign the $(i+1)$th cohort as follows:

a) If $\mathcal{H}$ is not empty, randomize the $(i+1)$th cohort to dose-schedules in $\mathcal{H}$ with the probability proportional to their desirability, given by

$$\lambda = \frac{\hat{U}(d_{j_k^*}, s_k)}{\sum_{\mathcal{H}} \hat{U}(d_{j_k^*+1}, s_k)}, \quad (d_{j_k^*}, s_k) \in \mathcal{H} \tag{3.16}$$

b) If $\mathcal{H}$ is empty, randomize the $(i+1)$th cohort to dose-schedules in $\mathcal{A}$ with the probability proportional to their desirability, given by

$$\lambda = \frac{\hat{U}(d_j, s_k)}{\sum_{\mathcal{A}} \hat{U}(d_j, s_k)}, \quad (d_j, s_k) \in \mathcal{A} \tag{3.17}$$

Step 3: Repeat Step 2 until the maximum sample size $N_{\max}$ is reached or the trial is early terminated due to toxicity and/or futility. In the case that the trial is not early terminated, select ODS as the dose-schedule with the highest utility in $\mathcal{A}$.

Similar to many optimization problems, a challenge of finding ODS is that the sequential dose-finding process is often stuck at local optimal, leading to low accuracy to identify the true optimal dose (Yuan et al., 2017). This issue is of particular concern for dose-schedule finding because the sequential searching process in a two-dimensional space is substantially more likely to be stuck at local optimal. We address this issue by two measurements. The first one is to prioritize Step 2(a) over 2(b) to encourage dose exploration, given that regimes are sufficiently safe as ensured by $\mathcal{H}$. The second measurement is to use adaptive randomization in Step 2(b) and 2(c) to provide additional freedom to explore the dose-schedule space, while accounting for patient benefit by assigning patients to more desirable regimes with higher probabilities. Yuan et al. (2017, Chapter 11) described the important role of adaptive randomization as an efficient tool to balance the exploration-versus-exploitation conflict in early phase clinical

trials. During the trial, we do not allow for skipping any untried dose in any schedule. When appropriate, equal randomization can also be used to simplify the logistics of trial conduct (Zhou et al., 2019).

3.4. Simulations

3.4.1. Simulation Setting

Following the advanced solid tumors trial, we considered five doses and two schedules, resulting in 10 dose-schedule regimes. The maximum sample $N_{\max} = 40$, treated in cohorts of size 2. The upper limit of the toxicity rate $\bar{p}_T = 0.3$, and the lower limit of the efficacy rate $\underline{p}_E = 0.3$. The utility elicited from clinicians is displayed in Table 1. We compared the PKIDS design with its counterpart that ignores the PK data. We denote the latter as the EffTox design to highlight that it makes the decision of patient allocation and ODS selection based on (Y_E, Y_T). In the EffTox design, marginal toxicity and efficacy probabilities are modeled as

$$\begin{aligned} \text{logit}(p_{E,jk}) &= \alpha_{0k} + \alpha_1 d_j + \gamma_1 s_k + \xi_1 (d_j \times s_k) \\ \text{logit}(p_{T,jk}) &= \beta_{0k} + \beta_1 d_j + \gamma_2 s_k + \xi_2 (d_j \times s_k) \end{aligned} \tag{3.18}$$

where the schedule effect is captured by schedule-specific intercepts α_{0k} and β_{0k}. The joint distribution of (Y_E, Y_T) follows the same Gumbel model as the PKIDS design. The EffTox design uses the same decision rule and dose-schedule finding algorithm as the PKIDS design to assign patients and determine the ODS. We choose the EffTox design for comparison because other existing designs define ODS with different decision rules, making the comparison difficult to interpret. Actually, the EffTox design may be viewed a modification of the design of Thall et al. (2007), by using the same decision rule as the PKIDS design, to facilitate a more meaningful comparison. We examined the operating characteristics of the PKIDS design, in comparison with the EffTox design,

under six scenarios with different locations of ODS, see Table 3.2. Figure 3.2 depicts the dose-toxicity, efficacy, and utility curves for the six scenarios. Under each scenario, we simulated 1000 trials.

3.4.2. Simulation Results

Table 3.3 and Table 3.4 provides the simulation results, including the selection percentage and the average numbers of patients treated at each dose-schedule based on 1000 simulated trials. Overall, PKIDS outperforms EffTox in most scenarios with a higher percentage of correct selection (PCS) of the ODS, and more patients allocated to the ODS.

In scenario 1, the ODS is (d_1, s_1). The PCS of PKIDS is 72.5%. In contrast, the PCS of EffTox is 61.0%. On average, PKID allocates 5.9 more patients to the ODS than EffTox. In scenario 2, the ODS is (d_3, s_1). PKIDS and EffTox has similar PCS, and the latter allocates slightly (1.6 on average) more patients to the ODS. Scenario 3 has (d_3, s_2) as the ODS. PKIDS outperforms EffTox with higher PCS (51.6% versus 32.0%) and more patients allocated to the ODS (9.8 versus 8.34). In scenario 4, the ODS is (d_4, s_2) with the highest utility of 53.81. This scenario is more challenging as one of its neighbor regimes (d_4, s_1) has a competitive utility of 49.47. The PCS of PKIDS is 56.0%, higher than that of EffTox (52.0%). In scenario 5, where (d_5, s_1) is the ODS, PKIDS outperforms EffTox with higher PCS (42.1% versus 6.0%). In scenario 6, the ODS is the highest dose-schedule (d_5, s_2). The PCS of PKIDS is 82.0%, whereas that of EffTox is 52.0%.

3.4.3. Sensitivity Analysis

We assessed the sensitivity of the PKIDS design to the specification of utility. We considered an alternative utility specification (see Table 3.4), which assigns a higher score of $\rho_{11} = 75$ to $(Y_E, Y_T) = (1, 1)$, compared to the utility used in the main simulation (see

Table 3.1). As shown in Figure 3.5, results are similar between the two specifications of the utility, suggesting that the PKIDS design is not sensitive to the specification of the utility. Such robustness is also observed in previous dose-finding studies using the utility approach (Guo and Yuan, 2017; Liu et al., 2018; Zhou et al., 2019; Lin et al., 2021). One main reason for such robustness is that the primary objective of using utility here is to rank the desirability of regimes for assigning patients and selecting ODS, not to focus on estimation and interpretation of the absolute value of the utility itself. As long as the rank is similar, which is generally robust to small estimation differences, the design will yield similar operating characteristics.

3.5. Discussion

We have proposed a Bayesian phase I-II trial design to optimize dose-schedule regimes based on the risk-benefit tradeoff. A main contribution of the proposed PKIDS design is to integrate PK data and modelling into dose-schedule optimization. This approach is not only well aligned with the biological and causal pathway underlying the effect of dose-schedule on toxicity and efficacy, but it also improves the accuracy of the dose-schedule finding as shown in the simulation study. The PKIDS design bridges dose-finding with pharmacology; two fields with limited interactions, despite being inherently related. There has been increasing interest and push to integrate these two fields for better decision making and drug optimization from both the clinical community (Ratain, 2014) and regulatory authorities (Mirat et al., 2021).

The PKIDS design can be extended in various ways. For example, we here consider binary toxicity and efficacy endpoints. Extension to continuous or time-to-event points are of interest for some trials based on these types of endpoints. In addition, PKIDS assumes that the endpoints are quickly observed to make adaptive dose-schedule assignment decisions. This assumption typically holds for PK endpoints, but may not

hold for toxicity and efficacy endpoints. Various methods for handling late-onset endpoints, e.g., Bayesian data augmentation (Liu et al., 2013; Jin et al., 2014), imputation (Yuan et al., 2018), weighting method (Cheung and Chappell, 2000), time-to-event approach (Yuan and Yin, 2009), and the approximated-likelihood approach (Lin and Yuan, 2019), can be used to address this issue. It also is of interest to incorporate patient characteristics (e.g., age, gender, and biomarkers) to PK, efficacy, and toxicity models to achieve precision medicine. This is more complicated than simply adding patient characteristics as covariates in the model. The implication is that by doing so, ODS will depend on individual patient characteristics. This demands a large sample size in order to find each patient's individual ODS. Of greater concern is the logistical challenge to determine how to implement the design in practice, and eventually deploy it to clinical use when the drug is effective.

Table 3.1: Utility for each possible toxicity and efficacy outcomes

	Efficacy	
Toxicity	Yes ($Y_E = 1$)	No ($Y_E = 0$)
No($Y_T = 0$)	100	35
Yes($Y_T = 1$)	60	0

Table 3.2: Six scenarios considered in simulation. Boldface indicates the efficacy probability, toxicity probability, and mean utility of the optimal dose-schedule (ODS) regime.

	Schedule					
Dose	$\Pr(Y_E = 1)$		$\Pr(Y_T = 1)$		**Utility**	
Scenario 1	1	2	1	2	1	2
1	**0.32**	0.34	**0.05**	0.17	**53.95**	50.79
2	0.38	0.43	0.27	0.34	49.63	50.19
3	0.47	0.50	0.50	0.70	46.73	41.13
4	0.53	0.56	0.80	0.87	39.24	38.45
5	0.58	0.60	0.93	0.95	37.42	37.87
Scenario 2						
1	0.13	0.17	0.06	0.10	41.30	42.44
2	0.22	0.28	0.16	0.21	43.47	45.48
3	**0.40**	0.43	**0.25**	0.40	**51.65**	47.95
4	0.46	0.48	0.50	0.60	46.11	43.62
5	0.50	0.52	0.70	0.80	41.13	38.63
Scenario 3						
1	0.05	0.10	0.10	0.12	34.72	37.22
2	0.15	0.21	0.15	0.20	39.35	41.38
3	0.28	**0.45**	0.24	**0.26**	44.38	**54.45**
4	0.48	0.52	0.50	0.65	47.36	44.23
5	0.55	0.57	0.80	0.90	40.46	37.93
Scenario 4						
1	0.05	0.10	0.01	0.02	37.90	40.79
2	0.15	0.17	0.04	0.07	43.31	43.52
3	0.22	0.28	0.10	0.14	45.65	48.04
4	0.36	**0.43**	0.24	**0.26**	49.47	**53.18**
5	0.44	0.47	0.50	0.55	44.86	44.87

Continued on next page

Table 3.2 – *Continued from previous page*

	Schedule					
Dose	$\Pr(Y_E = 1)$		$\Pr(Y_T = 1)$		**Utility**	
Scenario 5						
1	0.05	0.08	0.01	0.02	37.9	39.49
2	0.11	0.14	0.03	0.05	41.08	42.30
3	0.17	0.22	0.08	0.12	43.16	44.93
4	0.27	0.32	0.16	0.20	46.67	48.40
5	**0.45**	0.50	**0.24**	0.60	**55.21**	44.86
Scenario 6						
1	0.01	0.02	0.01	0.02	35.30	35.60
2	0.04	0.05	0.03	0.05	36.54	36.48
3	0.07	0.10	0.07	0.09	37.07	38.29
4	0.14	0.25	0.12	0.15	39.79	45.76
5	0.35	**0.45**	0.20	**0.25**	50.32	**54.83**

Table 3.3: Selection percentage and average percentage of patients treated at each dose-schedule. The optimal dose-schedule is in boldface.

	PKIDS				**EffTox**			
Dose	**Selection %**		**Average number of patients**		**Selection %**		**Average number of patients**	
Scenario 1	1	2	1	2	1	2	1	2
1	**72.5%**	11.4%	**20.2**	3.9	**61.0%**	19.0%	**14.3**	10.3
2	6.8%	5.9%	3.2	4.0	10.0%	1.0%	5.0	2.2
3	1.2%	0.4%	2.4	2.2	3.0%	2.0%	2.2	2.0
4	0.0%	0.0%	6.5	1.5	0.0%	0.0%	1.5	1.2
5	0.0%	0.9%	0.4	0.2	1.0%	0.0%	0.3	0.4
Scenario 2	1	2	1	2	1	2	1	2
1	8.0%	4.0%	4.3	2.9	0.0%	9.0%	2.0	5.8
2	7.0%	14.0%	3.9	5.4	0.0%	20.0%	2.1	5.9
3	**39.0%**	15.0%	**8.0**	5.0	**40.0%**	12.0%	**9.6**	5.3
4	2.0%	2.0%	3.1	2.4	5.0%	4.0%	2.1	2.7
5	0.0%	0.0%	1.8	1.4	2.0%	0.0%	1.1	1.6
Scenario 3	1	2	1	2	1	2	1	2
1	0.5%	0.3%	2.4	2.0	0.0%	7.0%	2.0	5.4
2	0.8%	1.1%	2.5	2.6	0.0%	22.0%	1.9	7.1
3	15.4%	**51.6%**	5.6	**9.8**	4.0%	**32.0%**	2.6	**8.3**
4	12.2%	4.1%	4.6	2.7	0.0%	0.0%	1.8	1.9
5	0.0%	0.0%	1.6	1.2	0.0%	0.0%	0.5	1.0
Scenario 4	1	2	1	2	1	2	1	2
1	1.0%	1.0%	2.1	2.0	0.0%	0.0%	2.0	2.4
2	0.0%	0.0%	2.0	2.1	0.0%	1.0%	2.0	2.4
3	2.0%	3.0%	2.7	3.3	0.0%	11.0%	2.1	4.9
4	12.0%	**56.0%**	5.5	**9.2**	7.0%	**52.0%**	3.2	**9.9**
5	17.0%	3.0%	6.9	3.4	1.0%	22.0%	2.6	8.1
Scenario 5	1	2	1	2	1	2	1	2
1	0.0%	0.1%	2.1	2.0	0.0%	1.0%	2.0	2.3
2	0.0%	0.0%	2.0	2.0	0.0%	1.0%	2.0	2.3

Continued on next page

Table 3.3 – *Continued from previous page*

	PKIDS				**EffTox**			
Dose	**Selection %**		**Average number of patients**		**Selection %**		**Average number of patients**	
3	0.0%	0.3%	2.1	2.2	0.0%	5.0%	2.0	4.2
4	1.9%	25.3%	3.0	5.8	6.0%	56.0%	2.9	10.3
5	**42.1%**	23.5%	**8.5**	9.3	**6.0%**	9.0%	**3.5**	6.5
Scenario 6	1	2	1	2	1	2	1	2
1	0.0%	0.0%	2.0	2.0	0.0%	0.0%	2.0	2.0
2	0.0%	0.0%	2.0	2.0	0.0%	0.0%	2.0	2.0
3	0.0%	0.0%	2.0	2.0	0.0%	0.0%	2.0	2.0
4	0.0%	0.0%	2.0	2.0	0.0%	1.0%	2.0	2.1
5	4.0%	**82.0%**	2.4	**18.9**	44.0%	**52.0%**	10.5	**13.0**

Table 3.4: Utility for each possible toxicity and efficacy outcomes

	Efficacy	
Toxicity	Yes ($Y_E = 1$)	No ($Y_E = 0$)
No($Y_T = 0$)	100	40
Yes($Y_T = 1$)	75	0

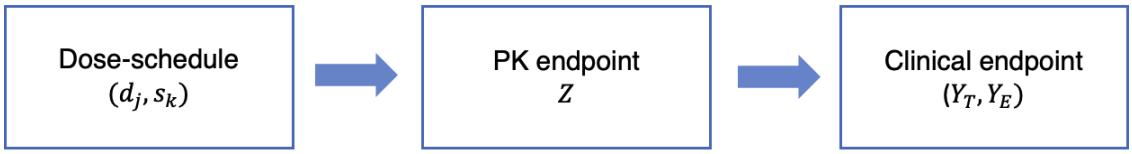

Figure 3.1: Casual path that the effects of dose-schedule pair (d_j, s_k) on clinical outcomes (Y_T, Y_E) are mediated by the plasma concentration Z of the drug.

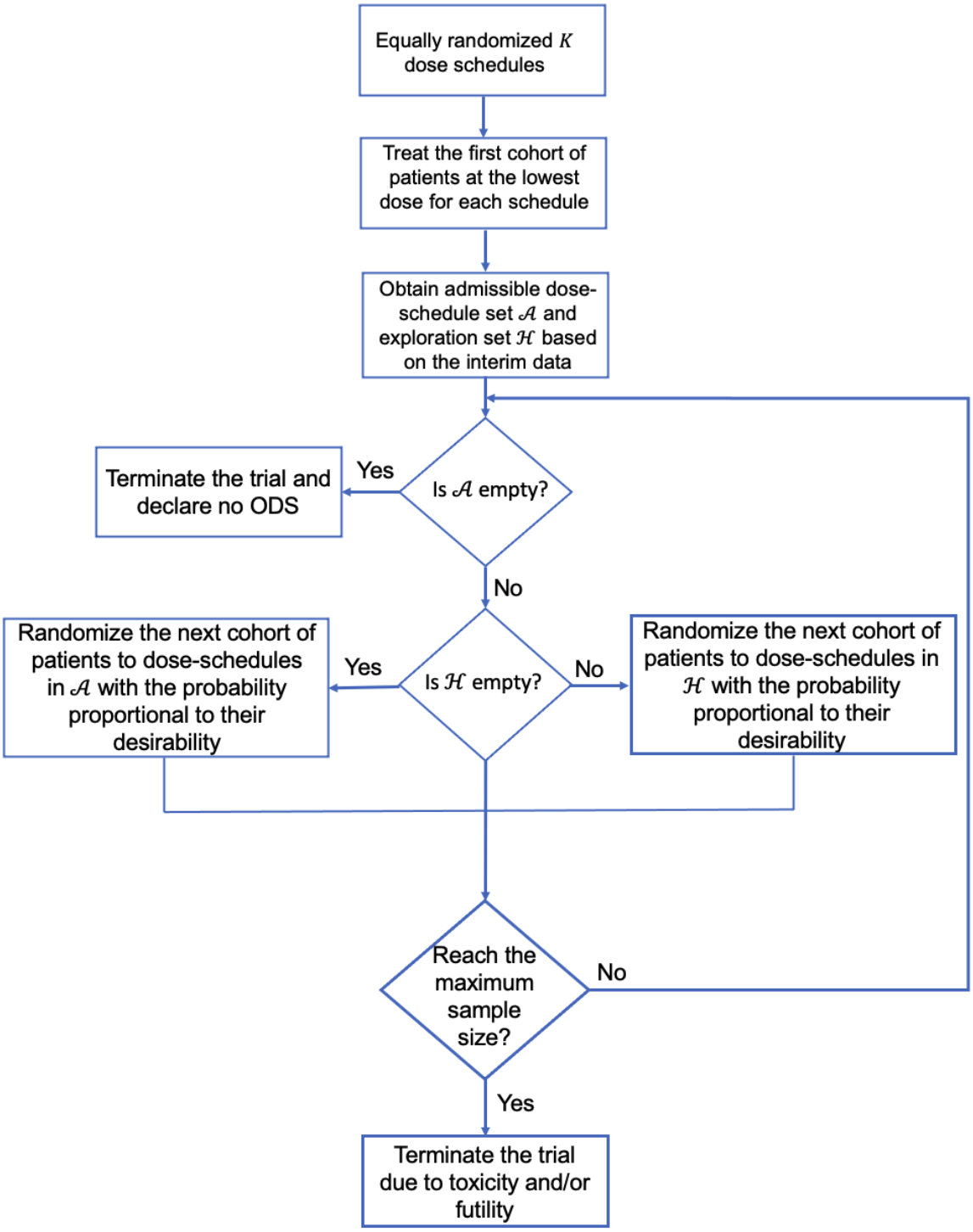

Figure 3.2: Dose finding algorithm of the PKIDS design

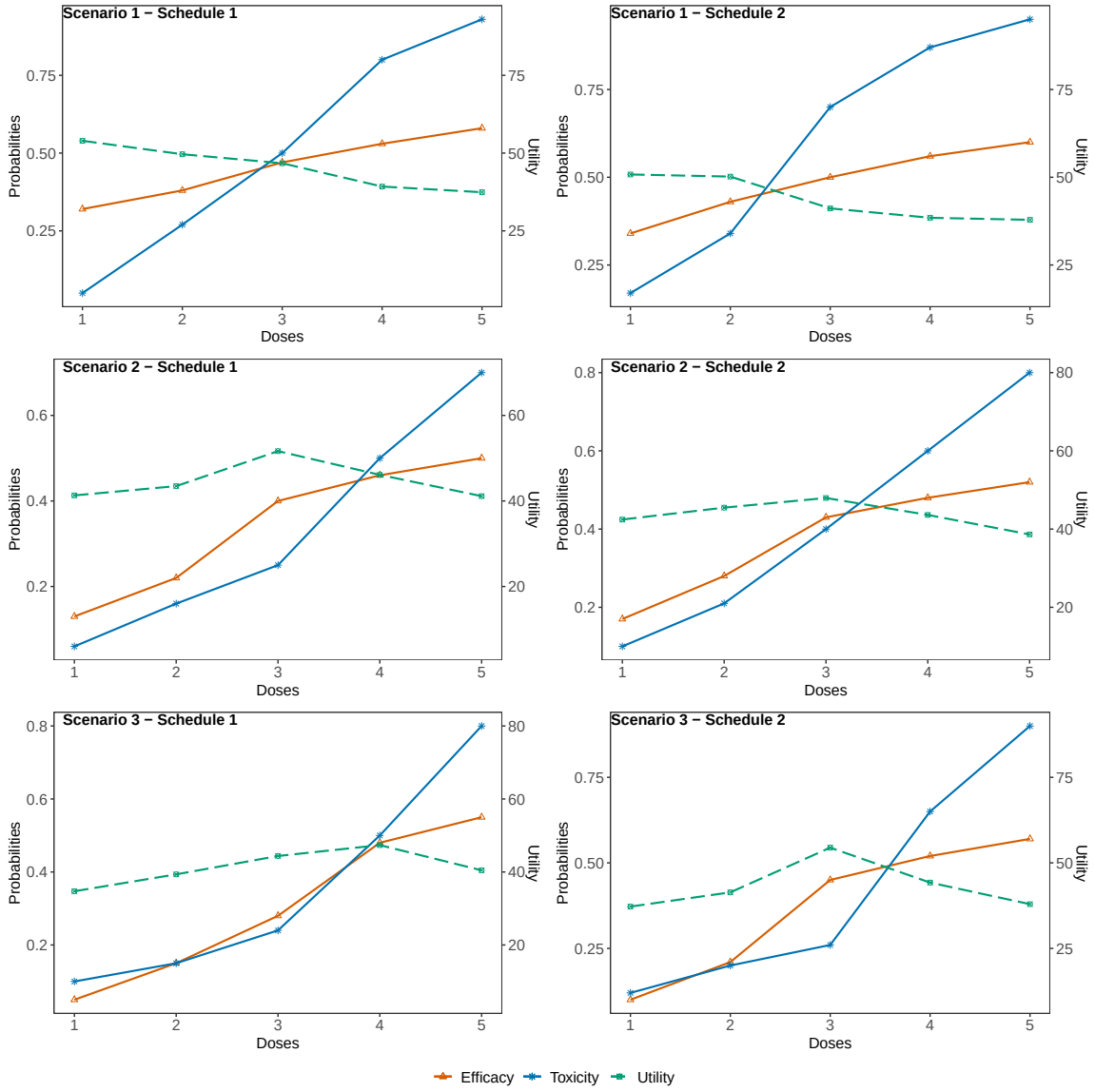

Figure 3.3: Dose-toxicity, efficacy, and utility curves under 1-3 scenarios.

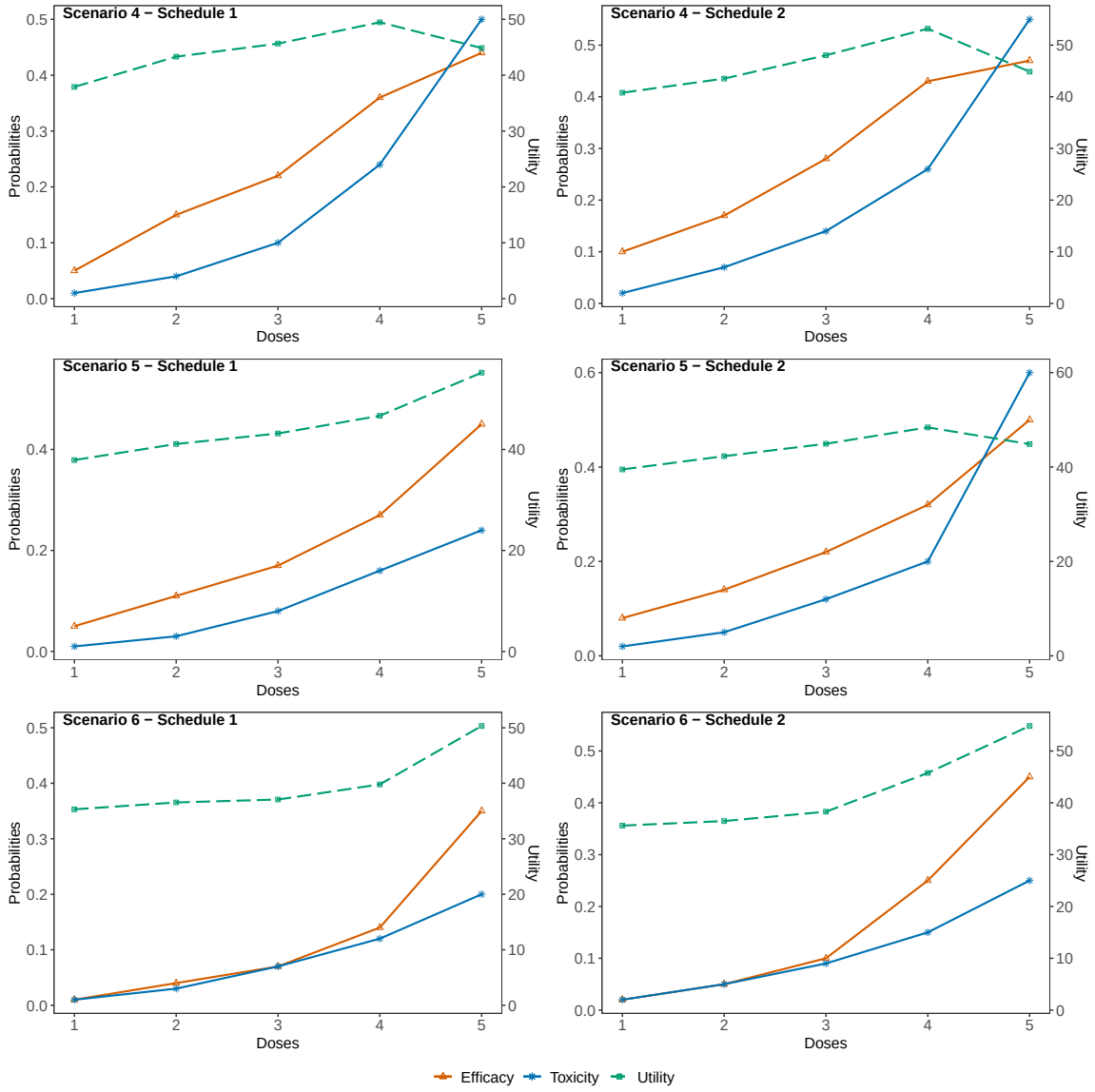

Figure 3.4: Dose-toxicity, efficacy, and utility curves under 4-6 scenarios.

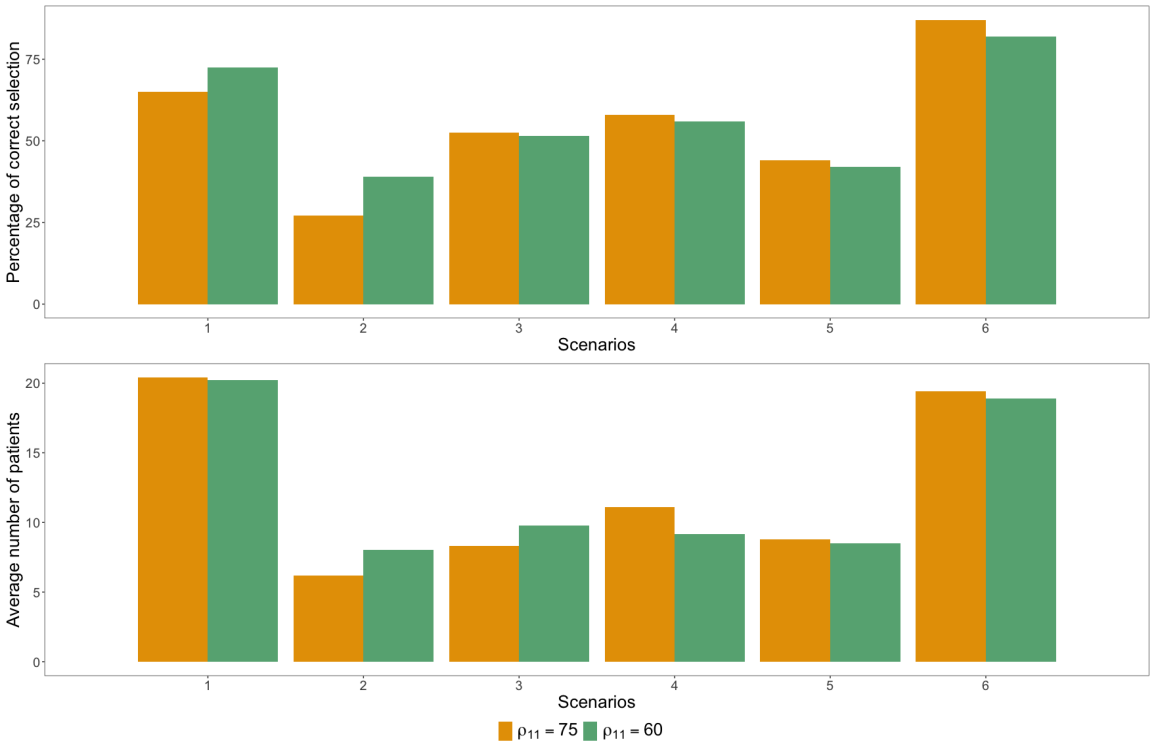

Figure 3.5: Sensitivity simulation results with an alternative utility.

CHAPTER 4

Bayesian Hierarchical Monitoring Design for Basket Trials

4.1. Introduction

With the rise in importance and interest in precision medicine, the basket trial has become a commonly-used trial design, a notable example is in phase II oncology trials (Wen et al., 2022; Friedman et al., 2022; Sherry et al., 2022; Subbiah et al., 2022; Fakih et al., 2022; Gupta et al., 2022; Bedard et al., 2022; How et al., 2021; Patel et al., 2021; Hahn et al., 2021). A basket trial investigates various patient populations with multiple tumor types treated under the same matched therapy; importantly, the cohorts (or baskets) in this design have a common biomarker signature (West, 2017; Friedman et al., 2015; Janiaud et al., 2019). Basket trials have the potential to increase the opportunities of registries for patients with rare cancers and potentially reduce the sample size within each basket relative to the equivalent collection of single histology trials by utilizing information across different cohorts. Additionally, simultaneous sub-studies of multiple tumor types may also accelerate drug development (Redig and Jänne, 2015; Berry, 2015; Garralda et al., 2019). However, conducting a basket trial is more resource-intensive than a trial of a single histology, as multiple cohorts are recruited and more patients are enrolled. Therefore, it is critical to develop innovative, yet reliable, basket trial designs to monitor drug efficacy and stop cohorts that fail to show evidence of activity as early as possible.

A particular area of interest in phase II clinical trials is efficacy monitoring, as about 75% of phase II trials fail to achieve a pre-specified minimum level of efficacy

(Thomas et al., 2021). Predetermined monitoring rules can avoid assigning a large number of patients to an inferior treatment. Several trial designs on efficacy monitoring have been proposed, however, most are only applicable to single-arm trials. The work by Simon (1989) proposed the first (frequentist) interim-monitoring design for phase II trials. Thall and Simon (1994) proposed a similar Bayesian design, which has inspired a series of extensions in the area of Bayesian phase II trials (see for example Thall et al. (1995); Cai et al. (2014); Zhou et al. (2017); Guo and Liu (2020)). However, these approaches are only applied in scenarios of single-arm trials, which is not applicable to basket trials.

The primary objective of basket trials is to evaluate if a novel treatment has sufficient efficacy in multiple baskets for further large confirmatory studies. The analysis strategy for basket trials is different from that of single-arm trials because we are investigating multiple histologies simultaneously. One strategy is pooled analysis combining all data across baskets, i.e. assuming homogeneity across baskets. However, this approach often fails when the treatment effect is heterogeneous across different cancer types. Alternatively, if the treatment effect is expected to be heterogeneous *a priori*, a stratified analysis can be performed separately in each basket. Due to limited sample size in each basket, this independent analysis often lacks sufficient power to detect the treatment effect. There are multiple possible methods for sharing information across baskets to potentially increase power. Cunanan et al. (2017) proposed an efficient frequentist design for phase II basket trials, which consider each basket as an independent Simon (1989) design, but combines baskets at the interim decision if two or more baskets appear to have homogeneous and beneficial response rates. Bayesian designs can share information across baskets using Bayesian model averaging (Simon et al., 2016; Psioda et al., 2021) or Bayesian Hierarchical Models (BHMs; Thall et al. (2003); Berry et al. (2013);

Neuenschwander et al. (2016); Liu et al. (2017); Chu and Yuan (2018)). However, we are interested in the setting with both a short-term endpoint and long-term endpoint, and none of the aforementioned work considers sharing information across baskets for trial designs with both short-term and long-term endpoints. This is important because it is possible that the short-term endpoint shows no effect from a novel treatment, while the long-term endpoint shows a meaningful improvement, however a candidate treatment for a particular histology may be ended early for futility due to the short-term outcome. Our proposed design is to use a BHM to effectively share information among baskets and use both short-term and long-term endpoints to yield valid results from interim-monitoring in phase II clinical trials.

Our research is motivated by the first FDA-approved histology-agnostic molecularly targeted therapy for Tropomyosin receptor kinases (TRK) fusion-positive cancers. TRK fusion was founded in 17 different cancers, including common cancers (breast, melanoma, lung cancer, etc.) and rare tumors such as secretory breast carcinoma and infantile fibrosarcoma (Vaishnavi et al., 2015; Drilon et al., 2018). The experimental drug Larotrectinib, a TRK inhibitor, has a high response rate in solid tumors that harbor TRK fusion and is well-tolerated among adults, adolescents, and pediatrics. Our motivating study includes three single-arm clinical trials: a phase I adult (LOXO-TRK-14001; NCT02122913), phase I-II children (SCOUT; NCT02637687), and a phase II basket trial involving adults and adolescents (NAVIGATE; NCT02576431) (Food and Association, 2018; Chen and Chi, 2018; Drilon et al., 2018). All of the aforementioned studies aim to investigate the efficacy and safety of Larotrectinib. The primary endpoint is the short-term endpoint, objective response rate (ORR), assessed by independent review using RECIST (Response Evaluation Criteria in Solid Tumors). The secondary (long-term) endpoint is the duration of response (DOR), defined as the time from the start date

of the initial response to the date of disease progression or death. The response rate was 81%, the median DOR was 35.2 months, and the longest DOR was 44.2 months (Hong et al., 2020). Therefore, target agents with long duration are worth further investigation. However, in most studies, investigators would terminate the trial due to prolonged DOR but no improvement in the response rate.

To overcome the limitations of current studies, we propose a BHM design for basket trials to monitor the efficacy of the investigational drug by incorporating both short-term endpoints (i.e., ORRs) and long-term endpoints (i.e., DORs). Our proposed design aims to avoid early terminations due to futility when the DOR is substantially prolonged with no improvement in the ORR. Two hierarchical levels are involved in the monitoring rule based on the short-term and long-term endpoints. Level II monitoring is triggered only if the stopping boundary for level I is reached. The trial will end due to futility if both levels reach their respective stopping boundaries.

Conditional on the latent variables, information borrowing can be conducted using a BHM. To our knowledge, no existing monitoring design for basket trials monitors both short-term and long-term endpoints. Simulation studies illustrate that our proposed design is easy to implement. And, under ideal conditions, our design can avoid early terminations for futility when an otherwise effective treatment involves a substantially prolonged DOR.

This chapter is organized as follows: Section 4.2 proposes the BHM design. Section 4.3 describes our proposed trial design. Section 4.4 presents simulation studies and the operating characteristics of the proposed design. Section 4.5 concludes.

4.2. Method

4.2.1. Objective response rate (ORR)

Consider patients sequentially entering a basket trial to investigate a targeted therapy in k different cancer types with matching tumor histology, biomarker, or gene expression. Objective response rate (ORR) is the primary endpoint and the basket trial aims to evaluate whether the ORR of the targeted therapy is clinically meaningful in each cancer type. This yields the following hypotheses:

$$H_{0k} : p_k \leq q_0 \quad \text{vs.} \quad H_{1k} : p_k \geq q_1 \quad \text{for} \quad k = 1, \ldots, K,$$

where p_k represents the objective response rate of the kth cancer type,$k = 1, \ldots, K$. q_0 denotes the unacceptable response rate and q_1 is the desired response rate, $q_1 > q_0$.

Assume patients receive a fixed follow-up period to evaluate the treatment responses and n_k patients with the kth cancer type are enrolled. X_k denotes the number of patients with responses in the kth cancer type, following a binomial distribution, $X_k \sim \text{Binomial}(p_k, n_k)$.

Let Z_k denote the latent cluster membership indicator, which will cluster cancer type k into groups with either high-response to treatment ($Z_k = 1$), or low response to treatment ($Z_k = 0$). Conditionally on Z_k, we assume that the treatment response rate p_k follows the following BHM,

$$\begin{gathered} \text{logit}(p_k)|Z_k \sim N(\mu_{0,Z_k}, \sigma^2_{0,Z_k}) \\ \mu_{0,Z_k=1} \sim N(\text{logit}(q_1), C^2_{01}), \quad \mu_{0,Z_k=0} \sim N(\text{logit}(q_0), C^2_{00}), \\ \sigma^2_{0,Z_k} \sim IG(a_0, b_0) \end{gathered} \tag{4.1}$$

where $N(\cdot)$ denotes a normal distribution and $IG(\cdot)$ denotes an inverse-gamma distribution. The parameters μ_{0,Z_k} and σ^2_{0,Z_k} represent the center of shrinkage and the borrowing strength respectively. Following our latent clustering, $\mu_{0,Z_k=1}$ is the center for high-response cluster, while $\mu_{0,Z_k=0}$ is the center for low-response cluster. Each of C^2_{00}, C^2_{01}, a_0 and b_0 are fixed hyperparameters that must be prespecified.

Because Z_k is a latent variable, and therefore never observed, the estimation of its value is joint with the other model parameters. We specify that Z_k follows a Bernoulli distribution

$$Z_k \sim \text{Ber}(\psi_k), \tag{4.2}$$

where ψ_k denotes the probability that the kth cancer type has a high response rate and the prior of ψ_k is set as $\text{Beta}(\alpha_0, \beta_0)$, where α_0 and β_0 are fixed hyperparameters.

4.2.2. Duration of response (DOR)

Let t_{ki} denote the duration of response for the ith monitored population (or responder) in the kth cancer type, and assume it follows an exponential distribution,

$$t_{ki} \sim Exp(\lambda_k). \tag{4.3}$$

A key innovation of our design is to include the effectiveness of DOR (the long-term response) in each cancer type. We assume

$$H_{0k} : \frac{\log 2}{\lambda_k} \leq r_0 \quad \text{vs.} \quad H_{1k} : \frac{\log 2}{\lambda_k} \geq r_1 \quad \text{for} \quad k = 1, \ldots, K,$$

where r_0 is an unacceptable level of median DOR and r_1 is a desired level of median DOR.

Similarly to our modeling of the short-term response ORR, let W_k denote the

latent cluster membership indicators, where $W_k = 1$ denotes that the kth cancer type has a long duration of response to the treatment and $W_k = 0$ denotes a short duration of response. Conditionally on W_k, we apply the following BHM to borrow information within the long and short DOR clusters, respectively.

$$\begin{gathered} \log(\lambda_k)|W_k \sim N(\mu_{1,W_k}, \sigma^2_{1,W_k}) \\ \mu_{1,W_k=1} \sim N\left(\log\frac{\log 2}{r_1}, C^2_{11}\right), \quad \mu_{1,W_k=0} \sim N\left(\log\frac{\log 2}{r_0}, C^2_{10}\right) \\ \sigma^2_{1,W_k} \sim IG(a_1, b_1) \end{gathered} \tag{4.4}$$

where $N(\cdot)$ denotes a normal distribution and $IG(\cdot)$ denotes an inverse-gamma distribution. As in the case of the model for the short-term outcome, μ_{1,Z_k} and σ^2_{1,Z_k} represent the center of shrinkage and the borrowing strength, and $\mu_{1,Z_k=1}$ is the center for high-response cluster while $\mu_{1,Z_k=0}$ is the center for low-response cluster. C^2_{10}, C^2_{11}, a_1 and b_1 are all fixed hyperparameters which must be prespecified.

For latent variable W_k, we assume that W_k follows a Bernoulli distribution:

$$W_k \sim Ber(\omega_k), \tag{4.5}$$

where parameter ω_k is the probability that the kth cancer type has a long duration of response and the prior of ω_k is set as Beta(α_1, β_1).

4.3. Trial design

In this section we develop Bayesian hierarchical monitoring rules for early stopping decision making. Up to the monitoring time, data from all enrolled patients are incorporated in level I monitoring, i.e. monitoring on the short-term outcome. Level II (the long-term outcome) monitoring is triggered only if the stopping rule of level I is

reached. The trial is terminated only if the stopping rules of both level I and level II monitoring are reached.

In the level I monitoring stage, the minimal clinical improvement δ is specified by clinicians based on historical information. If the response rate of the treatment is greater than $q_0 + \delta$, then the treatment is considered worth further investigation

$$Pr(p_k > q_0 + \delta | Data) > \varphi,$$

where the tuning parameter φ is the cut-off between 0 and 1. If the response rate satisfied the monitor criterion in level I, patients will continue to enroll without triggering level II monitoring. Otherwise, level II monitoring is executed for further evaluation.

In the level II monitoring stage, evaluation of DOR based on data from the respective responders is considered. Assuming m_k represents the number of responders in the kth group and N_k denotes the maximum sample size, the level 2 monitoring criterion is

$$Pr\left(\frac{\log 2}{\lambda} > r_0 | Data\right) > \left(\frac{m_k}{N_k}\right)^{\eta},$$

where the tuning parameter η is the cut-off between 0 and 1. If the level II monitoring rule is satisfied, patients can continue to enroll in the treatment. Otherwise, the trial is terminated for futility. In order to produce satisfactory operating characteristics, tuning parameters φ and η must be calibrated to control type I error rate and probability of early termination in a desired range.

4.4. Simulation

4.4.1. Simulation design

We conduct numerical simulation studies to evaluate the operating characteristics of the proposed BHM design for basket trials. Motivated by the clinical trials of Larotrectinib for TRK fusion–positive cancers described in the previous section, we set the primary endpoint as the ORR and the secondary endpoint as the DOR. Our main hypothesis is

$$H_{0k} : p_k \leq 0.5 \quad versus \quad H_{1k} : p_k \geq 0.7 \quad \text{for} \quad k = 1, \ldots, K,$$

where the null response rate is 0.5, and the response rate that is deemed promising is 0.7. As to the DOR of the responders, a median of 20 months offers only minimal benefit to patients, while a median of 35 months confers sufficient benefit to patients.

The interim analysis follows the same trial design pattern. Level II monitoring is triggered if the ORR of the kth cancer type reaches the level I monitoring stopping boundary. If the level II monitoring rule is satisfied, patient enrollment continues. Otherwise, the trial is terminated for futility. Let us assume that the investigation treatment involves a group of patients that includes five cancer types. The maximum sample size for each cancer type is 40, and the interim analysis is conducted when the first 20 patients complete the response assessment. The trial continues to enroll patients until all patients complete the response assessment. Some key hyperparameters are set as follows:

$$C_{01} = C_{00} = \frac{\text{logit}(q_1) - \text{logit}(q_0)}{6}, \tag{4.6}$$

$$C_{11} = C_{10} = \frac{\log(\log(2)/r_0) - \log(\log(2)/r_1)}{6}, \tag{4.7}$$

which means the priors for average treatment effects of two clusters are nearly non-overlapping. In terms of the variance of treatment effects within each cluster, σ^2_{0,Z_k} and σ^2_{1,W_k} follow an inverse-gamma distribution $IG(0.1, 0.1)$ given the limited sample size and number of subgroups. The hyperparameters ψ_k and ω_k are set to follow a non-informative prior $Beta(1, 1)$, which suggests an unbiased clustering of each cancer type.

We compare our proposed Bayesian hierarchical monitoring design with independent design and BHM design. The independent design models response rate of each cancer type independently by using a beta-binomial model for ORR, (i.e.,$p_k \sim beta(1, 1)$). A Gamma-Exponential model is used in the DOR with a conjugate prior $\lambda_k \sim Gamma(0.001, 0.001)$. The BHM design uses a traditional BHM model to borrow information directly from each cancer type.
For the ORR,

$$\begin{aligned} \text{logit}(p_k) &\sim N(\mu_0, \sigma_0^2) \\ \mu_0 &\sim N(\text{logit}(\frac{q_o + q_1}{2}), 10^{-6}) \\ \sigma_0^2 &\sim IG(0.1, 0.1) \end{aligned} \tag{4.8}$$

For the DOR,

$$\begin{aligned} log(\lambda_k) &\sim N(\mu_1, \sigma_1^2) \\ \mu_1 &\sim N(log\frac{log2}{(r_0 + r_1)/2}, 10^{-6}) \\ \sigma_1^2 &\sim IG(0.1, 0.1) \end{aligned} \tag{4.9}$$

We employed the same interim stopping rule in all the three designs to ensure they are comparable. The global null scenario is that the ORR and median DOR for all cancer types are 0.5 and 20 months, respectively. We calibrate the cutoffs δ and η of three

designs to ensure the type I error rate of about 0.05 for each cohort. The results of the calibrations are shown in Table 4.1.

Table 4.2 shows 12 different scenarios. In scenarios 1-4, the ORR and median DOR are the same across five cancer types. We established four conditions for true values of the ORR and the median DOR. These conditions comprise a full combination of an ORR that is either low or high and a DOR that is either short or prolonged. For Scenarios 5–8, the DOR remains constant, but the ORR is heterogeneous across cancer types. In these scenarios, the number of cohorts with high response rates decreases from 4 to 1, while the number of cohorts with low response rates increases from 1 to 4. The profile of the ORRs in Scenarios 9–12 is identical to their profile in Scenarios 5–8. The difference is that, at this time, cohorts with high response rates also exhibit longer median DORs. The results for each scenario are based on 10,000 simulations.

4.4.2. Simulation results

Figure 4.1 shows the percentage of rejecting H_{0k} of different cancer types under different scenarios. Table 4.3 summarizes simulation results. For cancer types with a true response rate of 0.5 or lower, the percentage of rejecting the null hypothesis equals the type I error rate. In addition, the probability of rejecting the null hypothesis corresponds to power when the true response rate is 0.7 or higher.

It is evident that regardless of the method is used, when the investigational drug does prolong the DOR, allowing for additional monitoring of the DORs of responders in the interim analysis can increase the power to detect a desirable treatment effect on the ORR. For instance, Scenario 4 versus Scenario 3, groups with high response rates in Scenarios 9–12 versus the corresponding groups in Scenarios 5–8. On the other hand, type I error rates are only slightly inflated when the ORR is undesirable, but the DOR is

prolonged, and this is still an acceptable level in phase II exploratory trials. For example, Scenario 2 versus Scenario 1, groups with low response rates in Scenarios 9–12 versus the corresponding groups in Scenarios 5–8. When the DOR is prolonged, the probability of terminating the trial decreases in the interim analysis, along with the expected sample size.

Compared to the independent design and the BHM design, our proposed design yields substantially more power to detect the treatment effect on the ORR when all five cohorts are homogeneous (Scenarios 3 and 4). Our proposed method performs at a level between the independent design and the BHM. When the cancer types are heterogeneous (Scenarios 5–12), our proposed design outperforms the BHM and independent designs. The BHM design fails to control the type I error rate. For instance, in Scenarios 5 and 9, the type I error rate is substantially inflated to over 25% for cancer type 1. In Scenarios 8 and 12, the BHM design has less power to detect the treatment effect. However, there is no serious type I error inflation or reduced power (Scenarios 8 and 12) in our proposed design because we consider two classification levels. The type I error rate that corresponds to cancer type 1 in Scenarios 5 and 9 falls between 10% and 15%, while other indication-specific type I error rates are kept below 10%.

4.4.3. Sensitivity Analysis

We also study the sensitivity of the BHM design with respect to the distribution of DORs. We evaluate the performance of the BHM design when the true distribution of DORs follows a Weibull or log-logistic distribution rather than the exponential distribution with a constant hazard shown in Section 4.2. We set the shape parameters for the Weibull distribution at 0.5 and 3 to generate decreasing and increasing hazards, respectively. For the log-logistic distribution, we set the shape parameter at 6 to generate a hump-shaped hazard. Figure 4.2 shows the hazard function of the exponential

distribution, two Weibull distributions and the log-logistic distribution when all median DORs are 20 months.

Figure4.3 shows the sensitivity analysis results based on the assumption that the shape parameters of the Weibull and log-logistic distributions are identical for all cancer types across all cancer types. Different true distributions of DORs produce different results. The use of exponential distributions in our proposed method could improve the probability of rejecting the null hypothesis when the true distribution of DORs follows a Weibull distribution with a shape parameter of 0.5. However, exponential distributions could reduce the probability of rejecting the null hypothesis when they follow a Weibull distribution with a shape parameter of 3 or a log-logistic distribution with a shape parameter of 6. The main reason for this difference is that the gaps between the median DOR and the mean DOR differ for those four distributions. For example, when the median DOR is 20, we set the means of the exponential distribution: Weibull distributions with shape parameters of 0.5 and 3; and log-logistic distributions of 28.85, 36.89, 20.18, and 5.24 months, respectively.

The use of exponential distributions in the model underestimates the hazard and leads to an overestimated median DOR when the true distribution of DORs follows a Weibull distribution with a shape parameter of 0.5. On the other hand, when the true distribution of DORs follows a log-logistic distribution, or follows a Weibull distribution with a shape parameter of 3, using the exponential distribution in the model will overestimate the hazard and lead to an underestimated median DOR.

According to the results of the sensitivity analysis, the use of exponential distributions is relatively robust in the scenarios we designed. The degree of type I error inflation and power reduction is limited. However, such results also reveal the risk that

our proposed method might misspecify the parameter model and thus render level II monitoring useless. Before practical clinical application, the appropriate parameter distribution must be carefully selected based on clinicians' previous clinical study data and advice.

4.5. Discussion

We propose a Bayesian hierarchical monitoring (BHM) design for basket trials to evaluate the treatment effects of novel immunotherapy and targeted therapy in cancer treatment. The main contribution of our proposed design is that it reduces the probability of early trial termination when substantial prolonged DORs arise. On the other hand, our proposed design enables investigators to evaluate treatment effects more comprehensively using multiple endpoints.

One innovation of our proposed design is that it incorporates short-term and long-term endpoints by evaluating the binary outcome (e.g., the ORR) in level I and by assessing the time-to-event outcome (e.g., the DOR) in level II. Level II monitoring is triggered if the stopping boundary in level I monitoring is reached. Otherwise, patient enrollment continues. The trial is terminated for futility if both level I and level II satisfy the stopping rules.

Another innovation of our proposed design is that it uses latent variables and the BHM approach to borrow information adaptively across cancer types (or baskets) in basket trials. In addition, information is sparse and limited at the beginning of the trial due to long follow-up periods. To overcome this challenge, we use a monotonic increase function for responders, so the trial is not terminated accidentally.

Simulation studies show that our proposed method has characteristics that far exceed the BHM and independent design. Our BHM design controls the type I error

rate more effectively, and it offers more power to detect treatment effects. It is well monitored, with both short- and long-term efficacy outcomes. It also accommodates toxicity outcomes, and it can be extended to encompass additional monitoring rules. Last, while our simulation is based on an oncological basket trial, but our design is also effective for use in basket trials for other diseases.

Table 4.1: Cutoffs of three designs

	Independent	BHM	Proposed
δ	0.898	0.868	0.891
η	0.030	0.035	0.028

Table 4.2: Scenarios of different cancer types

Scenario	Outcome	Cancer Type				
		1	2	3	4	5
1	ORR	0.5	0.5	0.5	0.5	0.5
	DOR	20	20	20	20	20
2	ORR	0.5	0.5	0.5	0.5	0.5
	DOR	35	35	35	35	35
3	ORR	0.7	0.7	0.7	0.7	0.7
	DOR	20	20	20	20	20
4	ORR	0.7	0.7	0.7	0.7	0.7
	DOR	35	35	35	35	35
5	ORR	0.5	0.7	0.7	0.7	0.7
	DOR	20	20	20	20	20
6	ORR	0.5	0.5	0.7	0.7	0.7
	DOR	20	20	20	20	20
7	ORR	0.5	0.5	0.5	0.7	0.7
	DOR	20	20	20	20	20
8	ORR	0.5	0.5	0.5	0.5	0.7
	DOR	20	20	20	20	20
9	ORR	0.5	0.7	0.7	0.7	0.7
	DOR	20	35	35	35	35
10	ORR	0.5	0.5	0.7	0.7	0.7
	DOR	20	20	35	35	35
11	ORR	0.5	0.5	0.5	0.7	0.7
	DOR	20	20	20	35	35
12	ORR	0.5	0.5	0.5	0.5	0.7
	DOR	20	20	20	20	35

Table 4.3: Simulation results of the BHM, independent, and Bayesian hierarchical monitoring designs

Scenario	Design		Results of different cancer types				
			1	2	3	4	5
1	Independent	% reject	5.0	5.0	5.0	4.9	5.1
		% stop	83.2	82.0	82.1	83.1	82.6
	BHM	% reject	4.9	5.3	5.2	4.8	4.8
		% stop	88.2	87.0	87.9	88.2	87.8
	Proposed	% reject	4.9	5.2	5.1	5.0	4.9
		% stop	86.9	86.0	86.6	87.1	86.7
2	Independent	% reject	6.5	6.5	6.5	6.3	6.5
		% stop	38.7	38.2	38.3	38.7	39.9
	BHM	% reject	8.8	8.8	8.9	8.3	8.5
		% stop	26.7	26.5	26.1	26.1	27.2
	Proposed	% reject	7.5	7.5	7.6	7.3	7.5
		% stop	36.2	36.2	36.0	36.5	37.6
3	Independent	% reject	74.6	73.9	74.6	74.8	73.4
		% stop	22.3	22.4	21.9	21.6	22.8
	BHM	% reject	93.6	93.0	93.5	93.8	93.4
		% stop	7.2	7.8	7.4	6.9	7.4
	Proposed	% reject	82.8	82.5	83.2	83.4	82.3
		% stop	18.8	19.0	18.4	17.9	19.4
4	Independent	% reject	82.7	82.1	82.6	82.5	81.8
		% stop	9.3	9.8	9.5	9.0	9.9
	BHM	% reject	97.6	97.2	97.4	97.5	97.2
		% stop	1.9	2.2	2.1	2.0	2.1
	Proposed	% reject	91.5	91.0	91.3	91.7	90.8
		% stop	6.8	7.2	7.0	6.6	7.3

Continued on next page

Table 4.3 – *Continued from previous page*

Scenario	Design		Results of different cancer types				
			1	2	3	4	5
5	Independent	% reject	4.8	73.7	74.7	74.7	73.9
		% stop	83.0	22.4	21.9	21.6	22.8
	BHM	% reject	25.8	88.6	89.2	89.3	88.6
		% stop	63.3	12.1	11.3	11.3	11.8
	Proposed	% reject	12.5	77.7	78.1	78.3	77.4
		% stop	80.2	21.6	21.5	21.0	22.3
6	Independent	% reject	4.8	5.0	74.6	75.8	73.6
		% stop	83.0	82.0	21.9	21.6	22.8
	BHM	% reject	18.8	18.2	83.7	84.0	83.7
		% stop	70.9	71.0	16.4	16.1	16.6
	Proposed	% reject	8.1	8.7	76.6	76.8	75.7
		% stop	83.4	82.7	22.1	21.8	22.9
7	Independent	% reject	4.8	5.1	5.0	74.8	73.9
		% stop	83.0	82.0	82.1	21.6	22.8
	BHM	% reject	13.3	13.6	13.4	77.3	76.3
		% stop	77.2	76.9	77.0	22.2	23.4
	Proposed	% reject	7	7.5	7.5	75.5	74.5
		% stop	84.3	83.3	83.5	23	24.2
8	Independent	% reject	4.6	5.2	4.8	4.5	73.6
		% stop	83.0	82.0	82.1	83.1	22.8
	BHM	% reject	8.4	9.1	9.4	8.9	65.7
		% stop	83.0	81.9	82.2	82.8	32.9
	Proposed	% reject	6.1	7.0	6.8	6.2	68.8
		% stop	85.0	83.9	84.2	85.1	29.1

Continued on next page

Table 4.3 – *Continued from previous page*

Scenario	Design		Results of different cancer types				
			1	2	3	4	5
9	Independent	% reject	4.7	82.1	82.4	82.9	82.3
		% stop	83.0	9.8	9.5	9.0	9.9
	BHM	% reject	27.9	94.7	94.9	95.2	94.8
		% stop	55.8	4.1	3.9	3.6	3.8
	Proposed	% reject	14.1	87.5	87.7	88.1	87.3
		% stop	77.7	8.9	8.8	8.3	8.9
10	Independent	% reject	4.9	5.0	82.4	83.0	82.3
		% stop	83.0	82.0	9.5	9.0	9.9
	BHM	% reject	20.3	19.7	91.4	92.0	91.8
		% stop	66.0	65.8	6.2	6.0	6.2
	Proposed	% reject	8.8	9.1	86.3	86.8	86.1
		% stop	81.6	80.5	9.4	9.0	9.7
11	Independent	% reject	4.7	5.0	5.1	82.9	82.1
		% stop	83.0	82.0	82.1	9.0	9.9
	BHM	% reject	14.1	14.6	14.3	87.7	86.9
		% stop	74.1	73.9	73.1	9.1	9.6
	Proposed	% reject	7.1	7.6	7.7	85.7	84.7
		% stop	83.0	81.9	82.0	9.9	10.9
12	Independent	% reject	4.7	5.3	4.6	4.6	82.1
		% stop	83.0	82.0	82.1	83.1	9.9
	BHM	% reject	9.1	9.7	10.0	9.5	78.6
		% stop	81.4	80.4	80.4	81.0	15.3
	Proposed	% reject	6.5	7.3	7.0	6.5	80.4
		% stop	84.2	83.2	83.4	84.4	13.4

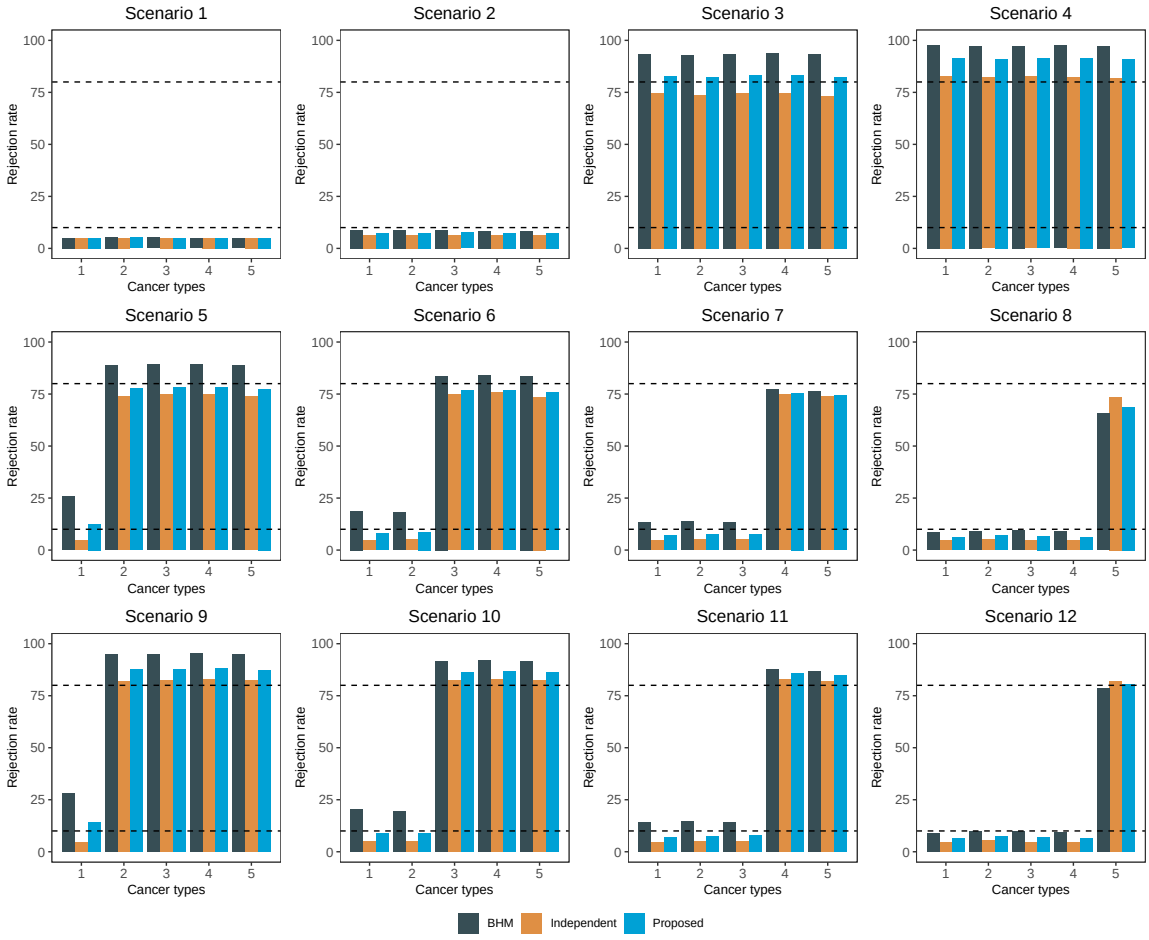

Figure 4.1: percentage of rejecting H_0 of different cancer types under different scenarios

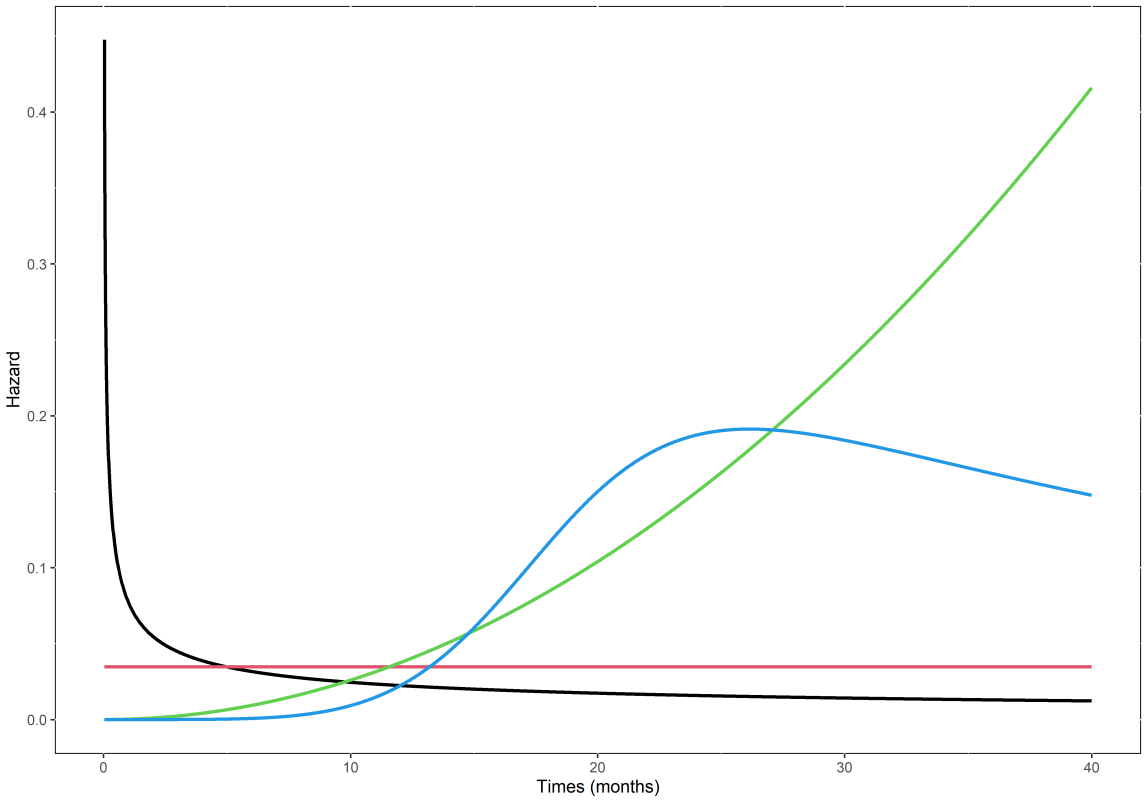

Figure 4.2: Hazard functions of DOR under the exponential (red curve), Weibull (black curve for decreasing hazard; green curve for increasing hazards), and log-logistic distributions (blue curve)

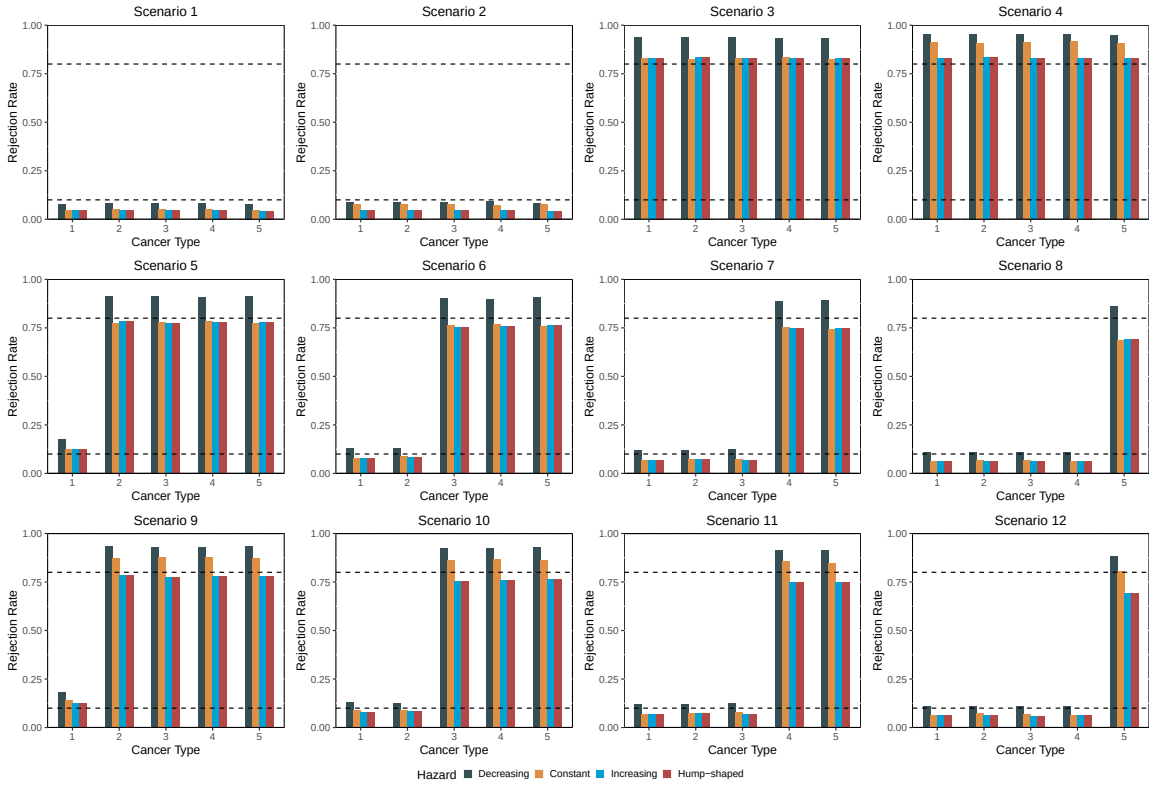

Figure 4.3: Results of the sensitivity analysis under 12 scenarios. The 4 bars from left to right represent the Weibull distribution with a decreasing hazard, the exponential distribution with a constant hazard, the Weibull distribution with an increasing hazard, and the log-logistic distribution with a hump-shaped hazard.

CHAPTER 5

Why There Are So Many Contradicted or Exaggerated Findings in Highly Cited Clinical Research?*

5.1. Introduction

Potentially exaggerated findings, and those contradicted in subsequent studies, are not unusual in clinical research (Benson and Hartz, 2000; Cappelleri et al., 1996; Ioannidis, 2005a,b; LeLorier et al., 1997). The uncertainties that arise are of particular concern when highly-cited clinical studies are involved, because of their great impact on clinical research and practice. Ioannidis (2005a) investigated 49 highly-cited original clinical research studies, each associated with more than 1000 citations, published from 1990 to 2003 in one of three major medical journals (New England Journal of Medicine, JAMA, or Lancet) or in a high-impact medical specialty journal. Of the 49 studies meeting the specified criteria, 45 had claimed positive findings—that the experimental intervention was effective. Findings of efficacy among 32% of those studies were contradicted in subsequent studies, or were shown to have potentially overestimated the efficacy of the experimental intervention. Another 44% of the studies had findings of efficacy that were later replicated, and 24% remained largely unchallenged during that publication period. Nonrandomized studies generally performed worse than randomized studies. The findings of a positive effect or the size of the effect reported in five out of six highly-cited nonrandomized studies were later contradicted or found to have been overestimated. However, randomized controlled trials (RCTs), generally considered the

*Chapter 5 is based upon "*Lu, M., Liu, S. and Yuan, Y. (2022) Why There Are So Many Contradicted or Exaggerated Findings in Highly Cited Clinical Research? Contemporary Clinical Trials.*", available online at: https://doi.org/10.1016/j.cct.2022.106782.

gold standard when evaluating the efficacy of clinical intervention, also suffered from a high percentage of contradiction in findings, with the published reports from 9 out of 39 RCTs contradicted by subsequent studies or found to have potentially overestimated the size of the effect. These results are of concern because of the rigorous standards often used to design and conduct such studies.

Given the widespread impact of highly-cited studies on clinical research, it is of great interest to understand the cause of such discrepancies. Ioannidis (2005a,b) identified characteristics that determine the probability of a research claim being true, including study power and bias, the number of other studies investigating the same question, and the ratio of true to no relationship found among the relationships probed in each scientific field. Specifically, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number of hypothesis testing; when there is greater flexibility in study designs; when there is a greater financial or other interest and related prejudice; and when more research teams are involved in a scientific field in chase of statistical significance.

The objective of this article is to conduct a statistical analysis of the highly-cited original clinical studies identified by Ioannidis (2005a) in order to understand the reasons for such high percentages of contradictory research findings. Highly-cited studies are often regarded as models or standards in the related research field, thus a careful analysis of such studies will provide insight into the process of clinical science and will help clinical researchers to appropriately evaluate and interpret experimental findings. Pinpointing the complete cause of contradictory or overestimated findings is a difficult challenge and would involve every component of clinical research, including study design, patient recruitment, trial conduct, data collection, data analysis, interpretation of the results, and publication of the findings. We will focus on the statistical analysis, particularly issues

related to the use of p values when testing the statistical hypothesis. Each of the highly-cited studies used p values to determine if the findings were statistically significant, i.e., if the p value was less than 0.05, the researchers claimed the experimental intervention to be superior to the control. In analyzing these studies, we demonstrated that p values strongly overstated the evidence of efficacy contained in the data: when p value equals 0.05, there is still 74.4% chance that the null hypothesis is true. This caused researchers to mistakenly conclude positive results (i.e., the false positive), which were then contradicted by subsequent large-scale studies that were statistically more powerful. For detailed theoretical perspectives of the pitfalls inherent in the use of p values in formal test procedures, see the work of Berger and Delampady (1987); Berger and Sellke (1987); Berger (2003); Goodman (1999a,b); Johnson (2013, 2019).

Our goal to statistically analyze the highly-cited studies in a meta-analysis is challenging. First, the data that can be retrieved from the published papers are very limited, as most studies only reported the sample size, effect size, p value and test statistics. Second, unlike traditional meta-analysis, which combines data from the same or similar scientific question, the highly-cited papers embraced a variety of medical questions and diseases. Such questions ranged from the effects of statins in the prevention of cardiovascular disease to tamoxifen for the prevention of breast cancer. Third, the researchers used different testing procedures to compare the new treatment with the standard treatment. To overcome these challenges, we based our analysis on test statistics, along the line of Johnson (2005) and Yuan and Johnson (2008). Test statistics can often be viewed as a condensed form of data and are reported in almost all studies. In addition, the distributions of test statistics are often known under the null hypothesis and are readily described under the alternative hypothesis. This greatly simplifies data modeling and converting p values to the posterior probability of hypotheses.

The rest of the paper is organized as follows. In section 5.2, we describe data extraction from the highly-cited clinical studies, and provide our analysis of the extracted data based on test statistics. Our findings appear in Section 5.3. We conclude with a brief discussion in Section 5.4.

5.2. Methods

5.2.1. Data Extraction

Ioannidis (2005a) identified 49 original highly-cited clinical studies according to the following criteria: each paper had received more than 1000 citations according to the index of the Institute for Scientific Information, and had been published between 1990 and 2003 in one of the three general medical journals with the highest impact (New England Journal of Medicine, JAMA, or Lancet) or in a medical specialty journal with an impact factor exceeding 7.0. (See Ioannidis (2005a), for further details.) Of the 49 original clinical research studies meeting the stated criteria, 42 were RCTs, 4 were prospective cohort studies, 2 were case series, and 1 was prospective case-control study (Note that Ioannidis (2005a) reported 43 RCTs, however, one of them (Ridker et al., 1997) actually was not RCT but a prospective case-control study). Since RCTs are considered the gold standard in evaluating the efficacy of clinical interventions, and are generally of better quality than observational studies, our analysis focused on the 42 highly-cited RCTs.

We abstracted data from the 42 highly-cited RCTs, including the sample size, number of events, p values, and the test statistic used to test the primary hypothesis (i.e., whether the experimental intervention was more effective than the control). We found that the p value for the primary hypothesis was often the first one that appeared in the abstract. Some papers provided the confidence interval and did not report the p values. In such cases, we calculated the p value based on the reported confidence

interval. Knowing the type of test statistic and associated p values, we could easily derive the value of the test statistics. For example, if a study employed the two-group logrank test to obtain a p value of 0.03, then the test statistic was χ^2 distributed with one degree of freedom under the null hypothesis, and its value was 4.71. The sample size we extracted was the "effective" sample size used in the primary hypothesis testing. This sample size might not have been the number of patients recruited in the study due to missing data or partial comparison. For example, the study ACTG019 (Volberding et al., 1990) compared low- and high-dose treatment groups with a control group. The main conclusion of the study was that the high dose of the treatment was effective. In that example, the effective sample size was the number of subjects in the high-dose group and the control group. During data extraction, we excluded three studies from our analysis due to unclear definitions of test statistics. We also excluded four RCTs with negative findings since these they actually refuted the findings of previous studies that had reported positive results (Ioannidis, 2005a), and thus may be treated as subsequent studies rather than original clinical research. Our final analysis was based on 35 highly-cited RCTs, which are listed in Table 5.1.

The 35 RCTs meeting our criteria all claimed positive findings. Of them, 25 had been followed by a larger study or by a published meta-analysis. Those 25 studies are listed under "contradicted or exaggerated studies" and "replicated studies" in Table 1 to indicate whether their findings agreed or disagreed with those of subsequent studies. The 10 highly-cited RCTs for which subsequent studies had not been reported are listed in Table 1 as "unchallenged studies." Subsequent RCTs usually have much larger sample sizes, and thus greater statistical power compared to original RCTs. Figure 5.1 depicts the relative sample sizes of original and subsequent studies. In 22 out of 25 subsequent studies, the samples sizes were at least three times as large as those of the original

studies. The smallest and the mean sample size of the subsequent studies were 2,440 and 21,444, respectively. Given such large sample sizes, the findings of subsequent studies may be reliable; therefore, we assume that their conclusions on the effectiveness of the experimental interventions were correct. Then the 35 RCTs can be divided into three groups according to the "true" status of the experimental interventions: 7 RCTs for which the experimental intervention was not more effective than the control, 18 RCTs for which the experimental intervention was more effective than the control, and 10 RCTs for which we did not know whether the experimental intervention or the control was more effective.

Most of the studies we analyzed used a time-to-event primary endpoint, and one of a variety of testing procedures (or test statistics) to test the efficacy of the experimental intervention. The majority of the RCTs used a logrank test or a Cox model to test the difference in hazards between participants given the experimental intervention and participants in the control group. Since the score of a test based on the Cox model is equivalent to that of a logrank test (Kleinbaum and Klein, 2012), we approximately treated tests based on the Cox model as logrank tests when deriving the distribution of the test statistics. The other RCTs ignored the time-to-event nature of the data , and treated the outcome as a binary variable in the form of a 2×2 table, then compared the percentage of events in the control group with that in the experimental intervention group using a binomial test, chi-squared test, or Fisher's exact test. Since all the studies had moderate-to-large sample sizes, such tests are approximately equivalent, and we refer to them as binomial-type tests. Hence, the test statistics used in the 35 RCTs can be roughly divided into two groups: 28 RCTs with logrank-type test statistics and 7 RCTs with binomial-type test statistics.

5.2.2. Model

Our statistical strategy was to model the test statistics. To proceed, we assumed the following experimental mechanism for generating test statistics for the highly-cited original RCTs:

(a) Physicians draw an experimental treatment i. With probability π, the treatment i is not different from the standard treatment (the null hypothesis H_0); with probability $1 - \pi$, the treatment i is different from the standard treatment (the alternative hypothesis H_1).

(b) An RCT is performed to generate a test statistic X_i or a p value.

(c) If the p value is less than 0.05, or equivalently $X_i > c_i$ where c_i is the critical value corresponding to a p value of 0.05, the H_0 is rejected, and results of the RCT are published.

The experimental model shown above is a highly simplified, hypothetic model. It is not necessarily correct, but may provide a reasonable approximation to the actual process of research and publication of highly-cited studies. Our experimental model assumes a simple publication selection model: only studies reaching statistical significance (i.e., p value < 0.05) are published. This censoring model was proposed by Lane and Dunlap (1978) and Hedges (1984) to correct the publication selection bias in meta-analysis. It has been further generalized to allow the conditional probability of selection to depend on the p value calculated for the study by Hedges (1992), Iyengar and Greenhouse (1988), and Dear and Begg (1992). The simple censoring selection model may not be appropriate for general publications since major journals do publish results from non-significant studies sometimes, but it is adequate for our purpose of modeling the highly-cited studies. As

noted by Ioannidis (2005a), highly-cited articles are a select sample with overrepesentation of positive findings since positive results are more likely to draw public attention and excite further scientific investigation and debate. In particular, the 35 RTCs we examined all reported significant results.

The primary endpoints of the 35 highly-cited RCTs were time-to-event in nature (e.g., time-to-remission, time-to-recurrence). For study i, let λ_{0i} denote the baseline hazard of the control group. A natural choice of hazard for the intervention group is of a proportional hazard form $\lambda_{1i} = e^{\theta_i}\lambda_{0i}$, where θ_i is the logarithm of the hazard ratio between the intervention and the control group. To account for the heterogeneity among RCTs due to different research subjects, study populations, and many other factors, we allowed study-specific effect size θ_i to vary across studies. The test of interest of the RCTs may be expressed as $H_0 : \theta_i = 0$ and $H_1 : \theta_i \neq 0$. We considered two-sided alternatives since all RCTs we investigated had conducted two-sided tests.

To convert test statistics (or p values) to posterior probabilities of H_0 (or Bayes factors), we first needed to derive the marginal distribution of test statistic X_i under H_0 and H_1. One advantage of modeling the test statistic is that its distribution is easily derived under H_0 and H_1. Under H_0, X_i usually follows a known distribution $f_{0i}(X_i)$, typically the standard normal distribution or a χ^2 distribution with a known degree of freedom, depending on the type of test statistic. Under H_1, X_i often follows a noncentral distribution in the same family as $f_{0i}(X_i)$. Let $p_{1i}(X_i|\delta(\theta_i))$ denote this noncentral distribution with a noncentrality parameter $\delta(\theta_i)$, which is a function of θ_i. Under the Bayesian framework, we needed to assign a prior distribution for the parameter θ_i and integrate it out to obtain the marginal distribution of X_i. Let $p(\theta_i|\tau)$ denote the prior density of θ_i, indexed by a parameter τ, then the marginal distribution of X_i under

H_1 is given by

$$f_{1i}(X_i|\tau) = \int p_{1i}(X_i|\delta(\theta_i))p(\theta_i|\tau)\,\mathrm{d}\theta_i. \tag{5.1}$$

As we discussed in Section 2.1, two classes of tests were used in the 35 RCTs to compare the efficacy of the experimental interventions with that of the controls. Logrank-type tests were used in 28 RCTs and binomial-type tests were used in seven RCTs. For the logrank test, the distribution of X_i under H_0 is χ^2_1, a χ^2 distribution with one degree of freedom. Under H_1, X_i follows a noncentral χ^2 distribution with one degree of freedom and the noncentrality parameter $\frac{1}{4}d_i\theta_i^2$, where d_i is the total number of events in the study. Following Johnson (2005), we assumed prior density of θ_i of the form,

$$p(\theta_i^2|\tau) \sim \tau\chi^2_1, \tag{5.2}$$

where τ is the scale parameter which determines both the location and dispersion of the effect size θ_i under the alternative hypothesis. Then it can be shown that the marginal distribution of X_i, $f_{1i}(X_i|\tau)$, is $(1 + \frac{1}{4}d_i\tau)\chi^2_1$.

Seven studies ignored the time-to-event nature of the data, and treated the outcome as a binary variable in the form of a 2×2 table. The percentages of events in the control group and the experimental intervention group were then compared using a binomial-type test. The binomial test statistic X_i^{bin} is often defined as

$$X_i^{bin} = \frac{\hat{p}_2 - \hat{p}_1}{\{[\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)]/n\}^{1/2}}$$

where $\hat{p}_1$ and $\hat{p}_2$ are observed percentages of events in the control group and the experimental intervention group. To convert the binomial test statistics to the same scale as the logrank test statistics, we defined $Y_i^{bin} = (X_i^{bin})^2$ and derived the marginal distributions

of Y_i^{bin} instead. Under H_0, Y_i^{bin} follows χ_1^2 since X_i^{bin} follows the standard normal distribution. Under H_1, it can be shown that Y_i^{bin} follows a noncentral chi-squared distribution with one degree of freedom and the noncentrality parameter

$$\delta(\theta_i) = \frac{n(p_1^{e^{\theta_i}} - p_1)^2}{p_1(1-p_1) + p_1^{e^{\theta_i}}(1 - p_1^{e^{\theta_i}})}$$

Applying the prior (5.2) and integrating out θ_i as (5.1), we obtained the marginal distribution of Y_i^{bin} under H_1. This marginal distribution does not have a closed form, but can be easily evaluated by numerical integration methods such as Gaussian quadratures. In practice, we do not know the value of p_1, but it can be estimated by $\hat{p}_1$.

With marginal distributions of X_i under H_0 and H_1 in hand, we derived the likelihood for the test statistics reported in each RCT. As described in the previous section, we divided the 35 highly-cited RCTs into three groups: 7 RCTs for which H_0 was true, 18 RCTs for which H_1 was true, and 10 RCTs for which we did not know the true status of H_0 and H_1. We denoted these three groups by $\mathcal{G}_0$, $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. Under the above experimental mechanism and assumptions, the likelihood contributed by the RCTs for which H_0 was true, i.e., $x_i \in \mathcal{G}_0$, is given by

$$\begin{aligned}
Pr(x_i, H_0|publish) &= \frac{Pr(x_i, H_0, publish)}{Pr(publish)} \\
&= \frac{Pr(H_0)Pr(x_i|H_0)Pr(publish|x_i, H_0)}{Pr(H_0)Pr(publish|H_0) + Pr(H_1)Pr(publish|H_1)} \\
&= \frac{\pi f_{0i}(x_i)Pr(x_i > c|x_i, H_0)}{Pr(H_0)Pr(X_i > c|H_0) + Pr(H_1)Pr(X_i > c|H_1)} \\
&= \frac{\pi f_{0i}(x_i)}{0.05\pi + (1-\pi)[1 - F_{1i}(c|\tau)]}
\end{aligned}$$

where $F_{1i}(\cdot)$ is the cumulative density function of a random variable with density $f_{1i}(\cdot)$. In a similar manner, it can be shown that the likelihood for the RCTs for which H_1 was

true, i.e., $x_i \in \mathcal{G}_1$, is given by

$$Pr(x_i, H_1|publish) = \frac{(1-\pi)f_{1i}(x_i|\tau)}{0.05\pi + (1-\pi)[1 - F_{1i}(c|\tau)]}$$

For studies that did not indicate the true status of the hypothesis (i.e., $x_i \in \mathcal{G}_2$), the likelihood was

$$\begin{aligned} Pr(x_i|publish) &= Pr(H_0)Pr(x_i|publish, H_0) + Pr(H_1)Pr(x_i|publish, H_1) \\ &= Pr(H_0)\frac{Pr(x_i|H_0)Pr(publish|x_i, H_0)}{Pr(publish|H_0)} + Pr(H_1)\frac{Pr(x_i|H_1)Pr(publish|x_i, H_1)}{Pr(publish|H_1)} \\ &= \pi\frac{f_{0i}(x_i)}{0.05} + (1-\pi)\frac{f_{1i}(x_i|\tau)}{1 - F_{1i}(c|\tau)}. \end{aligned}$$

Therefore, the likelihood of the test statistics $\boldsymbol{x} = \{x_1, \cdots, x_{35}\}$ from the 35 RCTs is given by

$$\begin{aligned} L(\boldsymbol{x}|publish) &= \prod_{x_i \in \mathcal{G}_0} Pr(x_i, H_0|publish) \prod_{x_i \in \mathcal{G}_1} Pr(x_i, H_1|publish) \prod_{x_i \in \mathcal{G}_2} Pr(x_i|publish) \\ &= \frac{\prod_{x_i \in \mathcal{G}_0} \pi f_{0i}(x_i) \prod_{x_i \in \mathcal{G}_1} (1-\pi)f_{1i}(x_i|\tau)}{\prod_{x_i \in \mathcal{G}_0, \mathcal{G}_1} [0.05\pi + (1-\pi)(1 - F_{1i}(x_i|\tau))]} \prod_{x_i \in \mathcal{G}_2} \left[\pi\frac{f_{0i}(x_i)}{0.05} + (1-\pi)\frac{f_{1i}(x_i|\tau)}{1 - F_{1i}(c|\tau)}\right] \end{aligned}$$

Based on this likelihood and assigning appropriate priors to parameters π and τ, we obtained the posterior estimates of π and τ using the Markov chain Monte Carlo method. Then, given a value of test statistics t, the probability that H_0 was true was obtained by Bayes' theorem, as follows:

$$\begin{aligned} Pr(H_0|t) &= \frac{Pr(H_0)Pr(t|H_0)}{Pr(H_0)Pr(t|H_0) + Pr(H_1)Pr(t|H_1)} \\ &= \frac{\pi f_0(t|\tau)}{\pi f_0(t) + (1-\pi)f_1(t|\tau)} \end{aligned} \tag{5.3}$$

5.2.3. Prior Specification and Estimation

We assigned independent noninformative prior distributions to the model parameters, as follows:

$$\begin{aligned} Pr(\tau) &\propto 1/\tau, \\ Pr(\pi) &\sim unif(0,1) \end{aligned}$$

The joint posterior distribution of $\{\tau, \pi\}$ is given by

$$Pr(\tau, \pi|\boldsymbol{x}) \propto Pr(\tau)Pr(\pi)L(\boldsymbol{x}|publish),$$

which can be sampled using the Metropolis-Hastings algorithm. Specifically, let $(\tau^{(t)}, \pi^{(t)})$ denote the tth posterior draws of the parameters. At the $t+1$ iteration, we generated candidate draws of τ and π, say (τ_t, π_t), from the following proposal densities,

$$\begin{aligned} Pr(\log(\tau_t)) &\sim N(\log(\tau^{(t)}), 0.7) \\ Pr(\text{logit}(\pi_t)) &\sim N(\text{logit}(\pi^{(t)}), 0.7) \end{aligned}$$

where $N(a, b)$ denotes a normal distribution with mean a and standard deviation b. The standard deviations of the proposal densities were chosen to yield reasonable Metropolis-Hastings jump distance and acceptance rate. The $(t+1)$th posterior draw $(\tau^{(t+1)}, \pi^{(t+1)})$ takes a value as follows,

$$(\tau^{(t+1)}, \pi^{(t+1)}) = \begin{cases} (\tau_t, \pi_t) & \text{with probability } \rho \\ (\tau^{(t)}, \pi^{(t)}) & \text{with probability } 1-\rho \end{cases}$$

where

$$\rho = \min \left\{ \frac{\pi_t(1-\pi_t)Pr(\tau_t, \pi_t|\boldsymbol{x})}{\pi^{(t)}(1-\pi^{(t)})Pr(\tau^{(t)}, \pi^{(t)}|\boldsymbol{x})}, 1 \right\}$$

We used 2000 iterations as burn in and 10000 iterations to obtain posterior draws. We monitored the convergence of the Markov chain by inspecting the trace plot.

5.3. Results

Table 1 shows p values, types of test statistics, and sample sizes for the 35 RCTs. We display the p values in Figure 5.2. An immediate observation was that contradicted or exaggerated studies tended to have larger p values than replicated studies. In particular, 4 out of 7 contradicted or exaggerated studies reported p values larger than 0.01, compared with 1 out of 18 for the replicated studies. This result is reasonable since in an informal sense, p values measure evidence against the null hypothesis. The smaller p values suggest more evidence of effectiveness. The replicated studies have stronger evidence of superiority of the experimental intervention than the contradicted or exaggerated studies. However, the problem is that, according to the current practice of using a p value of 0.05 as a measure of significance, the positive findings of 35 RCTs are all statistically significant, then why are such high percentages of these findings later contradicted or deemed to have been exaggerated? This naturally raises questions: Is the p value a sensible measure of evidence of the null hypothesis? Is 0.05 a reasonable cutoff for significance?

Actually, the p value greatly overstates the evidence contained in the data. Figure 5.3 shows trace plots and posterior distributions of τ and π, suggesting that the Markov chains in our analysis are well mixed and have reasonably converged. The posterior means for τ and π are 0.166 and 0.819, with 95% credible intervals (CI) of (0.074, 0.324) and (0.681, 0.911), respectively. By plugging the posterior draws of π and τ into (5.3),

we can obtain the posterior distribution of H_0 for any given value of the test statistic (or p value). Figure 5.4 depicts the relationship between p values and the probability that the H_0 is true. This plot is based on logrank tests with 388 events, the mean number of events for the 35 highly-cited RCTs. Surprisingly, when the p value is equal to 0.05, there is a 74.4% (95% CI = (0.560, 0.883)) chance that the null is true. If the p value equals 0.01, 0.001, and 0.0005, the chance that the null is true is 43.9%, 9.8% and 5.6%, respectively, with 95% CI = (0.264, 0.664), (0.050, 0.210), and (0.028, 0.126). Clearly, using a p value of 0.05 as the criterion for significance causes an excessive number of studies to mistakenly claim positive findings (i.e., a high false positive rate), which are then contradicted by subsequent large-scale studies. That is one of the statistical reasons why there are so many contradictory findings among the highly-cited studies.

However, we want to emphasize that the highly-cited RCTs is a highly selected sample from clinical research, and our hypothetic model is approximated and highly simplified. Our primary objective is to provide a preliminary explanation why there are so many contradicted results in highly cited studies. The results based on the highly-cited studies may not be directly applicable to general clinical research. For example, the high-cited RCTs we analyzed all reported significant studies, but medical journals also publish insignificant results sometimes. To model general clinical research, the model need to be extended to account for that fact. Nevertheless, Berger and Sellke (1987) showed that in general p values overstate the experimental evidence, and the actual evidence against a null can differ by an order of magnitude from the p value.

5.4. Discussion

Clinical trial design and conduct is a very complex process, and differences in various components of the process, such as differences in enrollment, eligibility criteria, clinical procedures, and many others, could cause discrepancies between the findings of

related studies. From a statistical point of view, our data analysis reveals an important reason for such discrepancies: p values strongly overstate the experimental evidence, and many findings reported as statistically significant based on p values are not significant at all. Using p values does not provide any protection for consumers of classic statistical testing methods. One common misconception about p values is that a p value of 0.05 represents a only 5% chance that the null is true, and thus stands for strong evidence of effectiveness. However, based on our analysis, for highly-cited studies, when the p value was equal to 0.05, there was a 74.4% chance that the null is true. Consequently, the rote use of a p value of 0.05 as the criterion of significance strongly overstates the evidence and may lead to serious consequences.

Our findings are consistent with the American Statistical Association (ASA) Statement on Statistical Significance and P-Values (Wasserstein and Lazar, 2016). Figure 5.4 shows that the p value indeed provides a measure of the compatibility of the data with a specified statistical model. As noted by ASA Statement, "the smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions." At the same time, our results also highlight that p values do not measure the probability that the studied hypothesis is true, and by itself, a p value does not provide a good measure of evidence regarding a model or hypothesis, as emphasized by the ASA Statement. For example, a p value near 0.05 offers only weak evidence against the null hypothesis. One major contribution of this paper is that our results provide real-world data evidence to support ASA Statement with high-impact clinical studies.

One simple way to improve the reproducibility of clinical studies (or to decrease

the false positive rate) may be to use a smaller p value, such as 0.001 or less to claim significant results, as advocated by Johnson (2019). Such a small p value can be achieved either under a large effect size with a moderate sample size or a moderate effect size with a large sample size. Given the reality of limited resources, many studies of small or moderate sample sizes do not afford enough sample size to detect a moderate effect size with such a small p value. Under such circumstances we should be cautious in our interpretation of the results. For example, we may choose to be conservative and retain certain reservations when accepting a conclusion with a p value of 0.01. The question then is, how conservative should we be? This exposes another pitfall of the p value: it does not provide a definite measure of the strength of evidence. In medical journals, we may conclude that a p value of 0.02 is (marginally) significant and 0.002 is highly significant, but these numbers do not tell us the difference between 0.02 and 0.002 quantitatively in terms of the evidence of effectiveness, i.e., a p value of 0.002 does not represent 10 times the level of evidence to support the alternative hypothesis compared to a p value of 0.02. (For further discussion of the use of p values in medical research, see, for example, the work of Goodman (1999a) and Goodman (1999b).)

A systematic and better approach to improve the reproducibility is to use the posterior probability of the hypotheses (the Bayes factor) to report experimental results, since it provides a more precise and objective measure of evidence contained in the data than p values. This will effectively decrease the false positive rate caused by the use of a p value of 0.05 as the criterion of significance. In addition, the posterior probability of the hypotheses has an intuitive interpretation, for example, a posterior probability of 0.3 for the alternative hypothesis means that there is a 30% chance that the alternative is true based on the current experiment, and a posterior probability of 0.6 represents two times that level of evidence. Wasserstein and Lazar (2016) provide more guidance and

discussion on the use of p value and how to improve the reproducibility of studies.

Table 5.1: Highly-Cited Randomized Clinical Trials

Studies (year published) [a]	Type of Intervention and Disease[a]	p Values	Tests	Effective Sample Sizes (Number of events)
Contradicted or exaggerated studies				
HA-1A Sepsis (1991)	Monoclonal antibody to endotoxin for gram-negative sepsis	1.4×10^{-2}	logrank	200(77)
CHAOS (1996)	Vitamin E to prevent MI and death in patients with CAD	5.0×10^{-3}	Cox	2002(105)
ACTG019 (1990)	Zidovudine in asymptomatic HIV-1 infection	2.0×10^{-3}	logrank	881(44)
PAMI (1993)	Angioplasty vs tPA thrombolysis in acute MI	2.0×10^{-2}	binomial	395(34)
STRESS (1994)	Stent vs balloon angioplasty in CAD	4.6×10^{-2}	chi-squared	336(123)
BENESTENT (1994)	Stent vs balloon angioplasty in single-vessel CAD	2.0×10^{-2}	chi-squared	516(128)
ACAS (1995)	Endarterectomy in asymptomatic stenosis 60%	4.0×10^{-3}	logrank	1659(85)
Replicated studies				
Moertel et al (1990)	Levamisole and fluorouracil for colon cancer	5.9×10^{-5}	logrank	619(258)
NASCET (1991)	Carotid endarterectomy in high-grade stenosis	1.2×10^{-6}	logrank	659(87)
SOLVD (1991)	Enalapril in patients with LV dysfunction	3.6×10^{-3}	logrank	2569(962)
SAVE (1992)	Captopril for patients after MI	1.9×10^{-2}	logrank	2231(503)
EPIC (1994)	7E3 in high-risk angioplasty	8.0×10^{-3}	logrank	1404(148)
WOSCOPS (1995)	Pravastatin in hypercholesterolemia	1.0×10^{-4}	logrank	6595(422)
CARE (1996)	Pravastatin after MI with average cholesterol	3.0×10^{-3}	logrank	4195(486)
US Carvedilol (1996)	Carvedilol for CHF	2.2×10^{-6}	Cox	1094(53)
ACTG320 (1997)	Triple therapy with indinavir vs 2 nucleosides in HIV-1 infection	1.0×10^{-3}	logrank	1156(96)
EPILOG (1997)	Abciximab glycoprotein IIb/IIIa blockade in PCI	1.8×10^{-6}	logrank	1874(157)
HIT (1998)	Interferon alfa-2b ribavirin vs interferon alone for chronic hepatitis C	1.4×10^{-9}	Fisher's exact	453(116)
LIPID (1998)	Pravastatin for secondary CAD prevention	3.8×10^{-4}	logrank	9014(660)
SHEP (1991)	Treatment of systolic hypertension in elderly adults	3.0×10^{-4}	logrank	4736(37)
AFCAPS/TexCAPS (1998)	Lovastatin for primary CAD prevention with average cholesterol	7.5×10^{-5}	logrank	6605(299)
4S (1994)	Simvastatin in hypercholesterolemia with previous CAD	3.0×10^{-4}	logrank	4444(438)
IHIT (1998)	Interferon alfa-2b ribavirin vs interferon alone for chronic hepatitis C	2.0×10^{-9}	Fisher's exact	555(172)
CIBIS-II (1999)	Bisoprolol for CHF	5.9×10^{-5}	Logrank	2647(384)
NSABP P-1 (1998)	Tamoxifen for breast cancer prevention	5.2×10^{-7}	binomial	13175(368)
Unchallenged studies				
V-HeFT II (1991)	Enalapril vs hydralazine isosorbide for CHF	1.6×10^{-2}	logrank	804(285)
Captopril Collaborative (1993)	Captopril for slowing disease progression in diabetic nephropathy	7.0×10^{-3}	logrank	409(68)
DCCT (1993)	Intensive management to reduce type 1 diabetes complications	1.8×10^{-9}	logrank	715(114)
ACTG076 (1994)	Zidovudine to reduce perinatal HIV-1 transmission	6.0×10^{-5}	logrank	363(53)
MRC Vitamin (1991)	Folate to prevent neural tube defects	5.0×10^{-3}	chi-squared	1195(27)
RALES (1999)	Spironolactone in severe CHF	7.6×10^{-6}	logrank	1663(670)
HOPE (2000)	Ramipril to prevent CAD in high-risk patients without LV dysfunction/CHF	2.2×10^{-6}	logrank	9297(1477)
CAPRIE (1996)	Clopidogrel vs aspirin in patients at risk of ischemic events	4.3×10^{-2}	logrank	19185(1960)
HOT (1998)	Intensive blood-pressure lowering/low-dose aspirin in hypertension	3.0×10^{-2}	Cox	18790(683)
UKPDS34 (1998)	Intensive management of type 2 diabetes with insulin or sulphonylureas	2.3×10^{-3}	logrank	753(256)

[a] Abbreviations and descriptions of the studies are adopted from Ioannidis (2005a).

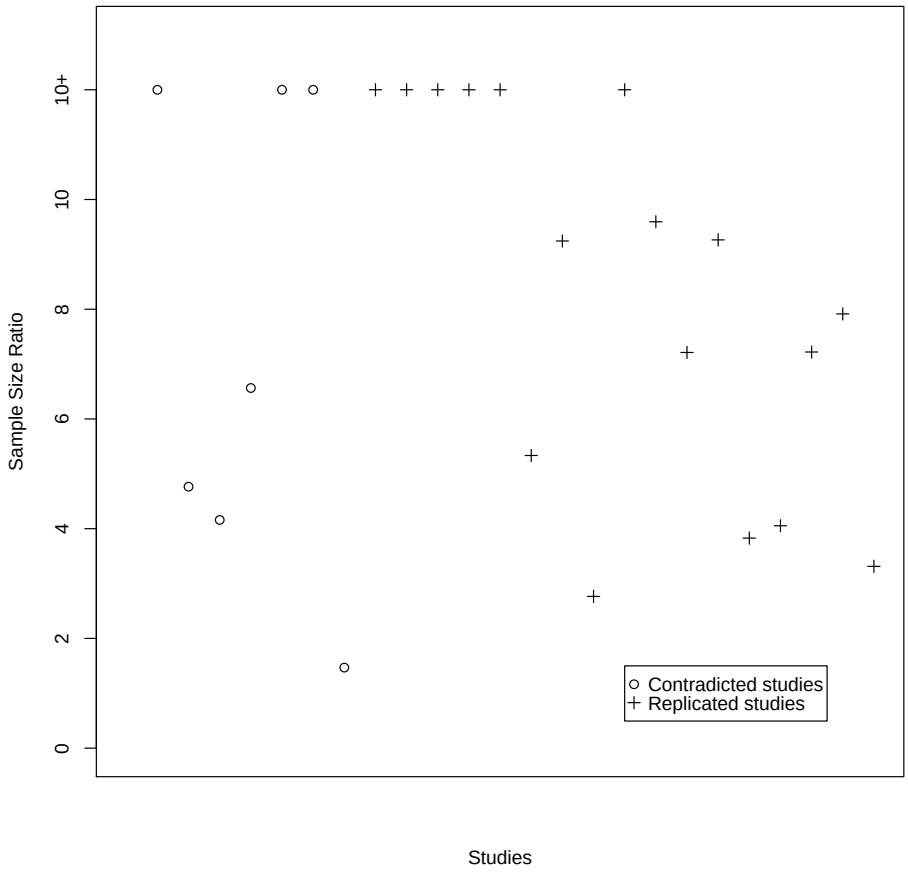

Figure 5.1: The ratio of sample sizes for 25 subsequent studies against corresponding original highly-cited controlled clinical trials. Crosses denote the contradicted studies, and circles denote the replicated studies.

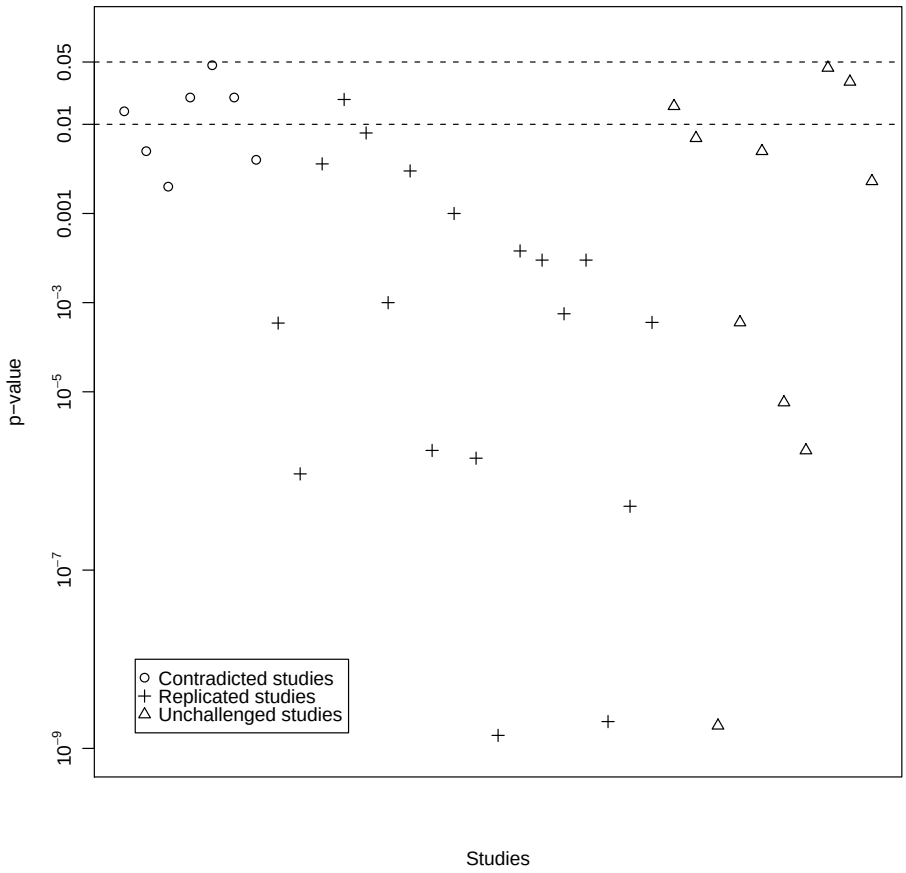

Figure 5.2: Logarithm of p values for 35 highly-cited controlled clinical trials, including 7 contradicted studies, 18 replicated studies and 10 unchallenged studies.

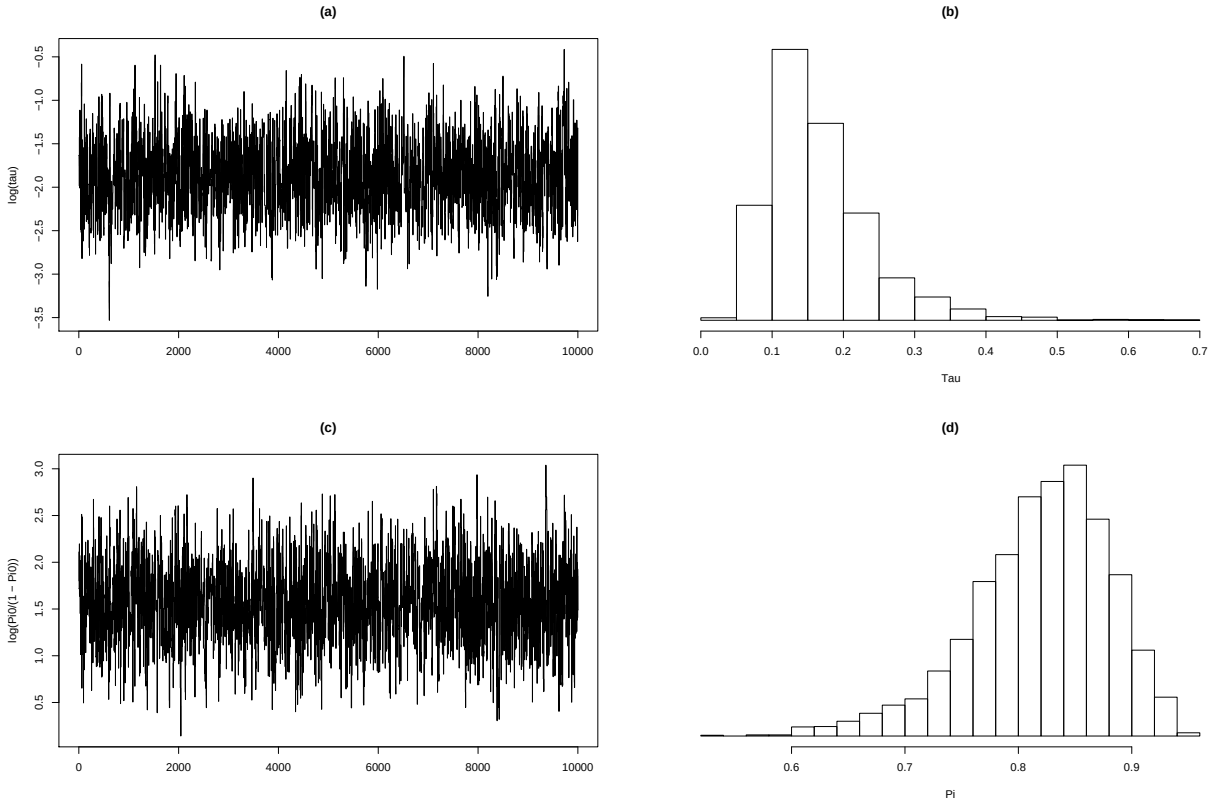

Figure 5.3: Analysis of the test statistics. (a) Trace plot of 10000 posterior draws of τ; (b) Posterior distribution of τ; (c) Trace plot of 10000 posterior draws of π; and (d) Posterior distribution of π.

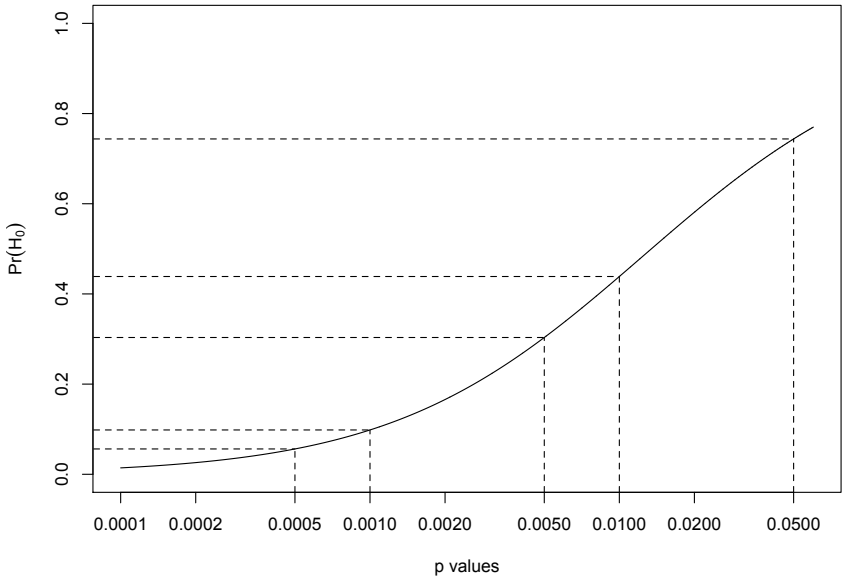

Figure 5.4: Probability that the null hypothesis is true under various p values, assuming the logrank test statistics and the 388 total number of events.

CHAPTER 6

Conclusion and future work

6.1. Conclusion

This dissertation focuses on early-phase adaptive clinical trial designs, especially under a Bayesian framework. We propose three novel, robust, and efficient Bayesian adaptive clinical trial designs that overcome the shortcomings of early-phase clinical trial designs in terms of dose-finding, dose-scheduling, and basket trial design.

Chapter 2 focuses on a seamless phase I/II dose-finding clinical trial design. The proposed Bayesian phase I/II trial design can efficiently identify the OBDC in a drug combination trial. When the toxicity or efficacy outcomes are late-onset, the time-to-event version of the design can utilize patients' follow-up data for decision making. We assign patients to the most appropriate dose combination by continuously updating the posterior distributions of toxicity and efficacy. When the toxicity or efficacy outcomes are late-onset, we propose a time-to-event version of the design that utilizes patients' follow-up data for decision making. Extensive simulation studies indicate that the proposed extensions of the BOIN12 are more straightforward to implement than the current phase I/II drug-combination designs. Various trial configurations show that the proposed designs also have outstanding operating characteristics for determining the OBDC.

Chapter 3 focuses on a phase I-II dose-schedule finding clinical trial design. A drug's administration schedule profoundly impacts its toxicity and efficacy profiles by changing its PK. PK is an inherent and indispensable component of dose-schedule optimization. The proposed Bayesian PK-integrated phase I-II design to optimize dose-schedule finding regimes identifies the optimal dose-schedule regime by integrating PK,

toxicity, and efficacy data. Based on the causal pathway through which dose and dose schedule affect PK, which in turn affects efficacy and toxicity, we model the three endpoints jointly by first specifying a Bayesian hierarchical model for the marginal distribution of the longitudinal dose-concentration process. Conditional on the drug concentration in plasma, we model toxicity and efficacy jointly as a function of the concentration. We quantify the risk–benefit tradeoff of regimes using utility (while continuously updating estimates of PK, toxicity, and effectiveness based on interim data), and we make adaptive decisions to assign new patients to appropriate dose-schedule regimes via adaptive randomization. The simulation study shows that the PKIDS design has desirable operating characteristics. It currently considers only binary toxicity and efficacy endpoints, but future studies could evaluate its extension to continuous or time-to-event endpoints. The PKIDS design also assumes that these endpoints are quickly observed.

Chapter 4 focuses on phase II basket trials with monitoring rules. The proposed Bayesian hierarchical monitoring design for basket trials reduces the probability of early termination when the DOR is substantially prolonged with no improvement in the response rate. The Bayesian hierarchical monitoring design includes two hierarchical levels: the short-term binary endpoint (ORR) and the long-term time-to-event endpoint (DOR). Conditional on latent subgroup indicators, we use the Bayesian hierarchical model to borrow information across different cancer types. Extensive simulation studies illustrate that our proposed design has outstanding operating characteristics compared to current trial designs. Our Bayesian hierarchical monitoring model yields higher power to detect treatment effects, and ideally, it can reduce the probability of early termination when the DOR is substantially prolonged with no improvements in the response rate.

Chapter 5 examines the reasons for the considerable number of contradicted or exaggerated findings in highly cited clinical research. Often, the outcomes of different

clinical studies of the same intervention contradict each other. We analyze a number of original, highly cited clinical studies that were later contradicted or were found to have overestimated the effects of experimental interventions by basing their analysis on test statistics within a Bayesian framework. We identify one source of contradictory results: the p values strongly overstated the experimental evidence. For highly cited studies, when the p value was .05, there was a 74.4% chance of confirming the null hypothesis. The use of a p value of .05 as the criterion for significance has caused many spurious positive findings that were later contradicted by large-scale studies.

Oncology therapies have developed dramatically as our knowledge of biomarkers and tumor biology increases. The concept of precision medicine is an important tool for defeating cancer methodically and systematically by focusing on matching the specific information of molecular mutation tumors to the most effective and accurate treatments. Compared to standard chemotherapeutic agents, novel therapeutic agents take longer to show clinical benefits such as tumor size reduction, transient pseudo progression, long-lasting partial response, or stable disease.

Clinical trial designs are critical in oncology drug development and treatment paradigms. Novel clinical trial designs accompany the introduction of precision medicine, and new treatment paradigms have rendered conventional dose-finding clinical trial designs inefficient and dysfunctional. By addressing the logistical and conceptual difficulties of existing clinical trial designs, our proposed design is also easy for clinicians to understand and implement. Our proposed novel methods perfectly embody the concept of precision medicine, which seeks to avoid a "one-size-fits-all" approach that ignores patient-specific characteristics.

6.2. Future work

Many unsettled issues and questions are still worth addressing in early-phase clinical trial designs.

In this dissertation, we only focus on novel therapies, such as targeted therapies and immunotherapies. However, an increasing number of treatments involving two or more therapies in cancer treatment are developed. For example, pembrolizumab plus chemotherapy for metastatic non–small-cell lung cancer (Gandhi et al., 2018) or advanced triple-negative breast cancer(Cortes et al., 2022). Results show that these novel therapies plus chemotherapies can significantly prolong the survival rate and provide a greater reduction of adverse events. Therefore, considering two or more treatments is one of the future directions for this dissertation.

R Shiny is a software platform commonly used to facilitate the use and understanding of adaptive clinical trial design methods. Many applications have been developed for clinicians and people who want to explore the performance of clinical trial designs but cannot understand the complicate statistically modeling. Facilitating, user-friendly, and easy-to-use Shiny web applications are another future work for this dissertation so that the users can explore the design by inputting the design parameters.

In chapter 2, our Comb-BOIN 12 design considers only two-drug combinations, but in actual practice, three or more drug combinations are also common in cancer treatment. One future research is to conduct three or more drug combination designs. One limitation of our design is that we only consider the homogeneity of patients. However, in actual practice, patients will have different sensitivities to the immune checkpoint inhibitors. Another interesting future direction is to assess patient heterogeneity.

Chapter 3 focuses on developing a dose-schedule finding design by integrating PK data. PK describes how the body does to the drug, whereas pharmacodynamic (PD) describes how the drug does to the body (Meibohm and Derendorf, 1997). Therefore, considering and integrating PD data for developing a dose-schedule finding design is one of the future directions of this study. Integrating PK/PD modeling and the differences between PK and PD data is also an interesting future direction of dose-schedule finding design. Here, in chapter 3, we assume each patient has PK data. However, in actual practice, examining PK/PD for each patient is costly and prolonged. Therefore, other covariates such as gender, age, and body mass index (BMI) can also be considered in future studies. Chemotherapies and immunotherapies have different mechanisms of action in the body. Since more cancer treatments focus on more than one therapy, It is also worth consideration in future studies.

Chapter 4 focused on short-term and long-term endpoints under novel cancer treatments, such as immunotherapies. A monitoring rule for combination therapies, such as immunotherapies plus chemotherapy, is worth considering in future studies. We only consider ORR as the primary endpoint and DOR as the secondary endpoint in chapter 4. However, in actual practice, ORR and DOR are not the only endpoints to assess the treatment effect. Monitoring multiple endpoints are also worth considering in future studies. In some trials, ORR and DOR are assumed to be equally crucial since short and long-term endpoints always correlate. An interesting future work includes jointly modeling and estimating the distribution of short and long-term endpoints in simultaneous monitoring. Here in chapter 4, we consider a phase II basket trial. However, the umbrella trial is also worth considering. An umbrella trial is also a novel adaptive trial design incorporating precision medicine into clinical trials, considering multiple treatment arms within one trial. Future open questions include monitoring rules for umbrella trials.

BIBLIOGRAPHY

Charu Aggarwal, Amy Prawira, Scott Antonia, Osama Rahma, Anthony Tolcher, Roger B Cohen, Yanyan Lou, Ralph Hauke, Nicholas Vogelzang, Dan P Zandberg, Arash R Kalebasty, Victoria Atkinson, Alex A Adjei, Mahesh Seetharam, Ariel Birnbaum, Andrew Weickhardt, Vinod Ganju, Anthony M Joshua, Rosetta Cavallo, Linda Peng, Xiaoyu Zhang, Sanjeev Kaul, Jan Baughman, Ezio Bonvini, Paul A Moore, Stacie M Goldbery, Fernanda I Arnaldez, Robert L Ferris, and Nehal J Lakhani. Dual checkpoint targeting of b7-h3 and pd-1 with enoblituzumab and pembrolizumab in advanced solid tumors: interim results from a multicenter phase i/ii trial. *Journal for immunotherapy of cancer*, 10(4), 2022.

Paolo A Ascierto, John M Kirkwood, Jean-Jacques Grob, Ester Simeone, Antonio M Grimaldi, Michele Maio, Giuseppe Palmieri, Alessandro Testori, Francesco M Marincola, and Nicola Mozzillo. The role of braf v600 mutation in melanoma. *Journal of translational medicine*, 10(1):1–9, 2012.

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.

Afsaneh Barzi, Nilofer Saba Azad, Yan Yang, Denice Tsao-Wei, Rabia Rehman, Marwan Fakih, Syma Iqbal, Anthony B El-Khoueiry, Joshua Millstein, Priya Jayachandran, Wu Zhang, and Heinz-Josef Lenz. Phase i/ii study of regorafenib (rego) and pembrolizumab (pembro) in refractory microsatellite stable colorectal cancer (msscrc)., 2022.

Philippe L Bedard, Shuli Li, Kari B Wisinski, Eddy S Yang, Sewanti A Limaye, Edith P Mitchell, James A Zwiebel, Jeffrey A Moscow, Robert J Gray, Victoria Wang, Lisa M McShane, Larry V Rubinstein, David R Patton, P Mickey Williams, Stanley R Hamilton, Barbara A Conley, Carlos L Arteaga, O'Dwyer Peter J Harris, Lyndsay N, Alice P Chen, and Keith T Flaherty. Phase ii study of afatinib in patients with tumors with human epidermal growth factor receptor 2–activating mutations: Results from the national cancer institute–molecular analysis for therapy choice ecog-acrin trial (eay131) subprotocol eay131-b. *JCO Precision Oncology*, 6:e2200165, 2022.

Kjell Benson and Arthur J Hartz. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, 342(25):1878–1886, 2000.

James O Berger. Could fisher, jeffreys and neyman have agreed on testing? *Statistical Science*, 18(1):1–32, 2003.

James O Berger and Mohan Delampady. Testing precise hypotheses. *Statistical Science*, pages 317–335, 1987.

James O Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American statistical Association*, 82(397):112–122, 1987.

Donald A Berry. The brave new world of clinical cancer research: Adaptive biomarker-driven trials integrating clinical practice with clinical research. *Molecular oncology*, 9 (5):951–959, 2015.

Scott M Berry, Kristine R Broglio, Susan Groshen, and Donald A Berry. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase ii oncology clinical trials. *Clinical Trials*, 10(5):720–734, 2013.

Carl Blomqvist, Inkeri Elomaa, Pentti Rissanen, Paivi Hietanen, Kaisu Nevasaari, and Leena Helle. Influence of treatment schedule on toxicity and efficacy of cyclophosphamide, epirubicin, and fluorouracil in metastatic breast cancer: a randomized trial comparing weekly and every-4-week administration. *Journal of clinical Oncology*, 11 (3):467–473, 1993.

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.

Thomas M Braun, Peter F Thall, Hoang Nguyen, and Marcos De Lima. Simultaneously optimizing dose and schedule of a new cytotoxic agent. *Clinical Trials*, 4(2):113–124, 2007.

Chunyan Cai, Ying Yuan, and Yuan Ji. A bayesian dose-finding design for oncology clinical trials of combinational biological agents. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 63(1):159, 2014.

Joseph C Cappelleri, PA John, Christopher H Schmid, Sarah D de Ferranti, Michael Aubert, Thomas C Chalmers, and Joseph Lau. Large trials vs meta-analysis of smaller trials: how do their results compare? *Jama*, 276(16):1332–1338, 1996.

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

Yu Chen and Ping Chi. Basket trial of trk inhibitors demonstrates efficacy in trk fusion-positive cancers. *Journal of hematology & oncology*, 11(1):1–5, 2018.

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

Dai Chihara, Ruitao Lin, Christopher R Flowers, Shanda R Finnigan, Lisa M Cordes, Yoko Fukuda, Erich P Huang, Larry V Rubinstein, Loretta J Nastoupil, S Percy Ivy, James H Doroshow, and Naoko Takebe. Early drug development in solid tumours: analysis of national cancer institute-sponsored phase 1 trials. *The Lancet*, 400(10351): 512–521, 2022.

Yiyi Chu and Ying Yuan. A bayesian basket trial design using a calibrated bayesian hierarchical model. *Clinical Trials*, 15(2):149–158, 2018.

Javier Cortes, Hope S Rugo, David W Cescon, Seock-Ah Im, Mastura M Yusof, Carlos Gallardo, Oleg Lipatov, Carlos H Barrios, Jose Perez-Garcia, Hiroji Iwata, Norikazu Masuda, Marco T Otero, Erhan Gokmen, Sherence Loi, Zifang Guo, Xuan Zhou, Vassiliki Karantza, Wilbur Pan, and Peter Schmid. Pembrolizumab plus chemotherapy in advanced triple-negative breast cancer. *New England Journal of Medicine*, 387(3): 217–226, 2022.

Kristen M Cunanan and Joseph S Koopmeiners. A bayesian adaptive phase i–ii trial design for optimizing the schedule of therapeutic cancer vaccines. *Statistics in medicine*, 36(1):43–53, 2017.

Kristen M Cunanan, Alexia Iasonos, Ronglai Shen, Colin B Begg, and Mithat Gönen. An efficient basket trial design. *Statistics in medicine*, 36(10):1568–1579, 2017.

Meindert Danhof, Gunnar Alvan, Svein G Dahl, Jochen Kuhlmann, and Gilles Paintaud. Mechanism-based pharmacokinetic–pharmacodynamic modeling—a new classification of biomarkers. *Pharmaceutical research*, 22(9):1432–1437, 2005.

Marie Davidian and A Ronald Gallant. Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. *Journal of Pharmacokinetics and Biopharmaceutics*, 20(5):529–556, 1992.

Keith BG Dear and Colin B Begg. An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, pages 237–245, 1992.

Wilfrid J Dixon and Alexander M Mood. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566, 1946.

Alexander Drilon, Theodore W Laetsch, Shivaani Kummar, Steven G DuBois, Ulrik N Lassen, George D Demetri, Michael Nathenson, Robert C Doebele, Anna F Farago, Al-

berto S Pappo, Brain Turpin, Afshin Dowlati, Marcia S Brose, Leo Mascarenhas, Noah Federman, Jordan Berlin, Wafik S El-Deiry, Christina Baik, John Deeken, Valentina Boni, Ramamoorthy Nagasubramanian, Mattew Taylor, Erin R Rudzinski, Funda Meric-Bernstam, Davendra PS Sohal, Patrick C Ma, Luis E Raez, Jaclyn F Hechtman, Ryma Benayed, Marc Ladanyi, Brain D Tuch, Kevin Ebata, Scott Cruickshank, Nora C Ku, Michael C Cox, Douglas S Hawkins, Davis S Hong, and David M Hyman. Efficacy of larotrectinib in trk fusion–positive cancers in adults and children. *New England Journal of Medicine*, 378(8):731–739, 2018.

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.

Marwan G Fakih, Scott Kopetz, Yasutoshi Kuboki, Tae Won Kim, Pamela N Munster, John C Krauss, Gerald S Falchook, Sae-Won Han, Volker Heinemann, Kei Muro, John H Strickler, David S Hong, Crystal S Henlinger, Gustavo Girotto, Myung-Ah Lee, Haby Henary, Qui Tran, Joseph K Park, Gataree Ngarmchamnanrith, Hans Prenen, and Timothy J Price DHthSc. Sotorasib for previously treated colorectal cancers with krasg12c mutation (codebreak100): a prespecified analysis of a single-arm, phase 2 trial. *The Lancet Oncology*, 23(1):115–124, 2022.

US Food and Drug Administration. Fda approves first-line immunotherapy for patients with msi-h/dmmr metastatic colorectal cancer, 2020.

US Food and Drug Association. Fda approves larotrectinib for solid tumors with ntrk gene fusions. *FDA. gov https://www. fda. gov/drugs/fda-approves-larotrectinib-solid-tumors-ntrk-gene-fusions-0*, 2018.

Adam A Friedman, Anthony Letai, David E Fisher, and Keith T Flaherty. Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer*, 15(12):747–756, 2015.

CF Friedman, AA D'Souza, AV Tinker, E Corral, V Gambardella, JW Goldman, S Loi, M Melisko, A Oaknin, I Spanggaard, AM VanderWalde, AL Frazier, B Zhang, LD Eli, and DB Solit. 559p neratinib in her2-mutant, recurrent/metastatic cervical cancer (r/m cc): Updated findings from the phase 2 summit basket trial. *Annals of Oncology*, 33:S803–S804, 2022.

Leena Gandhi, Delvys Rodríguez-Abreu, Shirish Gadgeel, Emilio Esteban, Enriqueta Felip, Flávia De Angelis, Manuel Domine, Philip Clingan, Maximilian J Hochmair, Steven F Powell, Susanna YS Cheng, Helge G Bischoff, Nir Peled, Francesco Grossi, Ross R Jennens, Martin Reck, Rina Hui, Edward B Garon, Michael Boyer, Belen

Rubio-Viqueira, Silvia Novello, Takayasu Kurata, Jhanelle E Grey, John Vida, Ziwen Wei, Jing Yang, Harry Raftopoulos, Catherine Pietanza, and Marina C Garassino. Pembrolizumab plus chemotherapy in metastatic non–small-cell lung cancer. *New England journal of medicine*, 378(22):2078–2092, 2018.

Tara C Gangadhar, Omid Hamid, David C Smith, Todd M Bauer, Jeffrey S Wasser, Jason J Luke, Ani S Balmanoukian, David R Kaufman, Yufan Zhao, Janet Maleski, Lance Leopold, and Thomas F Gajewski. Preliminary results from a phase i/ii study of epacadostat (incb024360) in combination with pembrolizumab in patients with selected advanced cancers. *Journal for immunotherapy of cancer*, 3(2):1–2, 2015.

Elena Garralda, Rodrigo Dienstmann, Alejandro Piris-Giménez, Irene Braña, Jordi Rodon, and Josep Tabernero. New clinical trial designs in the era of precision medicine. *Molecular oncology*, 13(3):549–557, 2019.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*, 2(4):1360–1383, 2008.

Wally R Gilks, Nicky G Best, and Keith KC Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(4):455–472, 1995.

CD Godwin, RP Gale, and RB Walter. Gemtuzumab ozogamicin in acute myeloid leukemia. *Leukemia*, 31(9):1855–1868, 2017.

Steven N Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of internal medicine*, 130(12):995–1004, 1999a.

Steven N Goodman. Toward evidence-based medical statistics. 2: The bayes factor. *Annals of internal medicine*, 130(12):1005–1013, 1999b.

Ted A Gooley, Paul J Martin, Lloyd D Fisher, and Mary Pettinger. Simulation as a design tool for phase i/ii clinical trials: an example from bone marrow transplantation. *Controlled Clinical Trials*, 15(6):450–462, 1994.

Burak Kürsad Günhan, Sebastian Weber, and Tim Friede. A bayesian time-to-event pharmacokinetic model for phase i dose-escalation trials with multiple schedules. *Statistics in Medicine*, 39(27):3986–4000, 2020.

Burak Kürsad Günhan, Sebastian Weber, Abdelkader Seroutou, and Tim Friede. Phase i dose-escalation oncology trials with sequential multiple schedules. *BMC Medical*

Research Methodology, 21(1):1–14, 2021.

Beibei Guo and Yisheng Li. Bayesian dose-finding designs for combination of molecularly targeted agents assuming partial stochastic ordering. *Statistics in medicine*, 34(5):859–875, 2015.

Beibei Guo and Suyu Liu. An optimal bayesian predictive probability design for phase ii clinical trials with simple and complicated endpoints. *Biometrical Journal*, 62(2): 339–349, 2020.

Beibei Guo and Ying Yuan. A comparative review of methods for comparing means using partially paired data. *Statistical methods in medical research*, 26(3):1323–1340, 2017.

Beibei Guo, Yisheng Li, and Ying Yuan. A dose–schedule finding design for phase i–ii clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(2):259–272, 2016.

Ranju Gupta, Funda Meric-Bernstam, Michael Rothe, Elizabeth Garrett-Mayer, Pam K Mangat, Stacy D'Andre, Eugene R Ahn, Raegan O'Lone, Susan Halabi, Gina N Grantham, and Richard L Schilsky. Pertuzumab plus trastuzumab in patients with colorectal cancer with erbb2 amplification or erbb2/3 mutations: Results from the tapur study. *JCO Precision Oncology*, 6:e2200306, 2022.

F Gyergyay, K Nagyvanyi, and I Bodrogi. Decreased toxicity schedule of suitinib in renal cell cancer: 2 weeks on/1 week off. *Journal of Clinical Oncology*, 27(15_suppl): e16113–e16113, 2009.

Andrew W Hahn, Jad Chahoud, Matthew T Campbell, Daniel D Karp, Jennifer Wang, Bettzy Stephen, Shi-Ming Tu, Curtis A Pettaway, and Aung Naing. Pembrolizumab for advanced penile cancer: a case series from a phase ii basket trial. *Investigational new drugs*, 39(5):1405–1410, 2021.

Omid Hamid, Todd Michael Bauer, Alexander I Spira, Anthony J Olszanski, Sandip Pravin Patel, Jeffrey S Wasser, David C Smith, Ani Sarkis Balmanoukian, Charu Aggarwal, Emmett V Schmidt, Yufan Zhao, Hema Gowda, and Tara C Gangadhar. Epacadostat plus pembrolizumab in patients with scchn: Preliminary phase i/ii results from echo-202/keynote-037., 2017.

John J Hanfelt, Rebecca S Slack, and Edmund A Gehan. A modification of simon's optimal design for phase ii trials when the criterion is median sample size. *Controlled Clinical Trials*, 20(6):555–566, 1999.

Larry V Hedges. Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, 9(1):61–85, 1984.

Larry V Hedges. Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2):246–255, 1992.

Ikuko Hirai, Keiji Tanese, Keitaro Fukuda, Takayuki Fusumae, Yoshio Nakamura, Yasunori Sato, Masayuki Amagai, and Takeru Funakoshi. Imatinib mesylate in combination with pembrolizumab in patients with advanced kit-mutant melanoma following progression on standard therapy: A phase i/ii trial and study protocol. *Medicine*, 100 (49), 2021.

David S Hong, Steven G DuBois, Shivaani Kummar, Anna F Farago, Catherine M Albert, Kristoffer S Rohrberg, Cornelis M van Tilburg, Ramamoorthy Nagasubramanian, Jordan D Berlin, Noah Federman, Leo Mascarenhas, Birgit Geoerger, Afshin Dowlati, Alberto S Pappo, Stefan Bielack, Francois Doz, Ray McDermott, Jyoti D Patel, Russell J Schilder, Makoto Tahara, Stefan M Pfister, Olaf Witt, Theodore W Laetsch, David M Hyman, and Alexander Drilon. Larotrectinib in patients with trk fusion-positive solid tumours: a pooled analysis of three phase 1/2 clinical trials. *The Lancet Oncology*, 21(4):531–540, 2020.

Nadine Houede, Peter F Thall, Hoang Nguyen, Xavier Paoletti, and Andrew Kramar. Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase i/ii trials. *Biometrics*, 66(2):532–540, 2010.

Jeffrey A How, Amir A Jazaeri, Pamela T Soliman, Nicole D Fleming, Jing Gong, Sarina A Piha-Paul, Filip Janku, Bettzy Stephen, and Aung Naing. Pembrolizumab in vaginal and vulvar squamous cell carcinoma: A case series from a phase ii basket trial. *Scientific reports*, 11(1):1–7, 2021.

John PA Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *Jama*, 294(2):218–228, 2005a.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2 (8):e124, 2005b.

Satish Iyengar and Joel B Greenhouse. Selection models and the file drawer problem. *Statistical Science*, pages 109–117, 1988.

Perrine Janiaud, Stylianos Serghiou, and John PA Ioannidis. New clinical trial designs in the era of precision medicine: an overview of definitions, strengths, weaknesses, and

current use in oncology. *Cancer treatment reviews*, 73:20–30, 2019.

Emily Y Jen, Chia-Wen Ko, Jee Eun Lee, Pedro L Del Valle, Antonina Aydanian, Charles Jewell, Kelly J Norsworthy, Donna Przepiorka, Lei Nie, and Jiang Liu. Fda approval: gemtuzumab ozogamicin for the treatment of adults with newly diagnosed cd33-positive acute myeloid leukemia. *Clinical cancer research*, 24(14):3242–3246, 2018.

Yuan Ji, Ping Liu, Yisheng Li, and B Nebiyou Bekele. A modified toxicity probability interval method for dose-finding trials. *Clinical trials*, 7(6):653–663, 2010.

José L Jiménez, Sungjin Kim, and Mourad Tighiouart. A bayesian seamless phase i–ii trial design with two stages for cancer clinical trials with drug combinations. *Biometrical Journal*, 62(5):1300–1314, 2020.

Ick Hoon Jin, Suyu Liu, Peter F Thall, and Ying Yuan. Using data augmentation to facilitate conduct of phase i–ii clinical trials with delayed outcomes. *Journal of the American Statistical Association*, 109(506):525–536, 2014.

Daniel H Johnson, Salah E Bentebibel, Srisuda Lecagoonporn, Chantale Bernatchez, Cara L Haymaker, Ravi Murthy, Alda Tam, Cassian Yee, Rodabe Navroze Amaria, Sapna Pradyuman Patel, Hussein Abdul-Hassan Tawbi, Isabella Claudia Glitza, Michael A Davies, Wen-jen Hwu, Patrick Hwu, Willem W Overwijk, and Adi Diab. Phase i/ii dose escalation and expansion cohort safety and efficacy study of image guided intratumoral cd40 agonistic monoclonal antibody apx005m in combination with systemic pembrolizumab for treatment naive metastatic melanoma., 2018.

Valen E Johnson. Bayes factors based on test statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):689–701, 2005.

Valen E Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313–19317, 2013.

Valen E Johnson. Raise the bar rather than retire significance. *Nature*, 567(7749):461–462, 2019.

Valen E Johnson and John D Cook. Bayesian design of single-arm phase ii clinical trials with continuous monitoring. *Clinical Trials*, 6(3):217–226, 2009.

HM Jones and Karen Rowland-Yeo. Basic concepts in physiologically based pharmacokinetic modeling in drug discovery and development. *CPT: pharmacometrics & systems pharmacology*, 2(8):1–12, 2013.

Sin-Ho Jung, Mark Carey, and Kyung Mann Kim. Graphical search for two-stage designs for phase ii clinical trials. *Controlled clinical trials*, 22(4):367–372, 2001.

Akihito Kawazoe, Yasutoshi Kuboki, Eiji Shinozaki, Hiroki Hara, Tomohiro Nishina, Yoshito Komatsu, Satoshi Yuki, Masashi Wakabayashi, Shogo Nomura, Akihiro Sato, Takeshi Kuwata, Masahito Kawazu, Hiroyuki Mano, Yosuke Togashi, Hiroyoshi Nishikawa, and Takayuki Yoshino. Multicenter phase i/ii trial of napabucasin and pembrolizumab in patients with metastatic colorectal cancer (epoc1503/scoop trial). *Clinical Cancer Research*, 26(22):5887–5894, 2020.

Razelle Kurzrock, Chia-Chi Lin, Tsung-Che Wu, Brian P Hobbs, Roberto Carmagnani Pestana, and David S Hong. Moving beyond 3+ 3: the future of clinical trial design. *American Society of Clinical Oncology Educational Book*, 41:e133–e144, 2021.

David M Lane and William P Dunlap. Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, 31(2):107–112, 1978.

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

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

Kwang-Pyo Lee, Joo-Hyeon Lee, Tae-Shin Kim, Tack-Hoon Kim, Hee-Dong Park, Jin-Seok Byun, Min-Chul Kim, Won-Il Jeong, Diego F Calvisi, Jin-Man Kim, and Dae-Sik Lim. The hippo–salvador pathway restrains hepatic oval cell proliferation, liver size, and liver tumorigenesis. *Proceedings of the National Academy of Sciences*, 107(18): 8248–8253, 2010.

Jacques LeLorier, Genevieve Gregoire, Abdeltif Benhaddad, Julie Lapierre, and François Derderian. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine*, 337(8):536–542, 1997.

Yisheng Li, B Nebiyou Bekele, Yuan Ji, and John D Cook. Dose–schedule finding in phase i/ii clinical trials using a bayesian isotonic transformation. *Statistics in medicine*, 27 (24):4895–4913, 2008.

Ruitao Lin and Ying Yuan. On the relative efficiency of model-assisted designs: a conditional approach. *Journal of biopharmaceutical statistics*, 29(4):648–662, 2019.

Ruitao Lin, Yanhong Zhou, Fangrong Yan, Daniel Li, and Ying Yuan. Boin12: Bayesian optimal interval phase i/ii trial design for utility-based dose finding in immunotherapy and targeted therapies. *JCO precision oncology*, 4:1393–1402, 2020.

Ruitao Lin, Guosheng Yin, and Haolun Shi. Bayesian adaptive model selection design for optimal biological dose finding in phase i/ii clinical trials. *Biostatistics*, 2021.

Xiaolei Lin and Yuan Ji. The joint i3+ 3 (ji3+ 3) design for phase i/ii adoptive cell therapy clinical trials. *Journal of Biopharmaceutical Statistics*, 30(6):993–1005, 2020.

Yong Lin and Weichung J Shih. Adaptive two-stage designs for single-arm phase iia cancer clinical trials. *Biometrics*, 60(2):482–490, 2004.

Rong Liu, Zheyu Liu, Mercedeh Ghadessi, and Richardus Vonk. Increasing the efficiency of oncology basket trials using a bayesian approach. *Contemporary Clinical Trials*, 63: 67–72, 2017.

Suyu Liu and Valen E Johnson. A robust bayesian dose-finding design for phase i/ii clinical trials. *Biostatistics*, 17(2):249–263, 2016.

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.

Suyu Liu, Guosheng Yin, and Ying Yuan. Bayesian data augmentation dose finding with continual reassessment method and delayed toxicity. *The annals of applied statistics*, 7(4):1837, 2013.

Suyu Liu, Beibei Guo, and Ying Yuan. A bayesian phase i/ii trial design for immunotherapy. *Journal of the American Statistical Association*, 113(523):1016–1027, 2018.

Anthony R Mato, Jakub Svoboda, Eline T Luning Prak, Stephen J Schuster, Patricia Tsao, Colleen Dorsey, Pamela S Becker, Danielle M Brander, Sunita Dwivedy Nasta, Daniel J Landsburg, Cara M King, Beth Morrigan, Kaitlin Kennard, Lindsey E Roeker, Andrew D Zelenetz, Michelle Purdom, Dana Paskalis, Peter Sportelli, Hari P Miskin, Michael S Weiss, and Mazyar Shadman. Phase i/ii study of umbralisib (tgr-1202) in combination with ublituximab (tg-1101) and pembrolizumab in patients with relapsed/refractory cll and richter's transformation. *Blood*, 132:297, 2018.

Bernd Meibohm and H Derendorf. Basic concepts of pharmacokinetic/pharmacodynamic (pk/pd) modelling. *International journal of clinical pharmacology and therapeutics*, 35 (10):401–413, 1997.

Tara C Mitchell, Omid Hamid, David C Smith, Todd M Bauer, Jeffrey S Wasser, Anthony J Olszanski, Jason J Luke, Ani S Balmanoukian, Emmett V Schmidt, Yufan Zhao, Xiaohua Gong, Janet Maleski, Lance Leopold, and Thomas Gajewski. Epacadostat plus pembrolizumab in patients with advanced solid tumors: phase i results from a multicenter, open-label phase i/ii trial (echo-202/keynote-037). *Journal of Clinical Oncology*, 36(32):3223, 2018.

Reza Bayat Mokhtari, Tina S Homayouni, Narges Baluch, Evgeniya Morgatskaya, Sushil Kumar, Bikul Das, and Herman Yeger. Combination therapy in combating cancer. *Oncotarget*, 8(23):38022, 2017.

Robert J Motzer, Thomas E Hutson, Mark R Olsen, Gary R Hudes, John M Burke, William J Edenfield, George Wilding, Neeraj Agarwal, John A Thompson, David Cella, Akintunde Bello, Beata Korytowsky, Jingyu Yuan, Olga Valota, Bridget Martell, Subramanian Hariharan, and Robert A Figlin. Randomized phase ii trial of sunitinib on an intermittent versus continuous dosing schedule as first-line therapy for advanced renal cell carcinoma. *Journal of clinical oncology*, 30(12):1371–1377, 2012.

Rongji Mu, Ying Yuan, Jin Xu, Sumithra J Mandrekar, and Jun Yin. gboin: a unified model-assisted phase i trial design accounting for toxicity grades, and binary or continuous end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(2):289–308, 2019.

Thomas A Murray, Ying Yuan, Peter F Thall, Joan H Elizondo, and Wayne L Hofstetter. A utility-based design for randomized comparative trials with ordinal outcomes and prognostic subgroups. *Biometrics*, 74(3):1095–1103, 2018.

Beat Neuenschwander, Alessandro Matano, Zhongwen Tang, Satrajit Roychoudhury, Simon Wandel, and SA Bailey. Bayesian industry approach to phase i combination trials in oncology. *Statistical methods in drug combination studies*, 2015:95–135, 2015.

Beat Neuenschwander, Simon Wandel, Satrajit Roychoudhury, and Stuart Bailey. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical statistics*, 15(2):123–134, 2016.

Sandip Pravin Patel, Edward Mayerson, Young Kwang Chae, Jonathan Strosberg, Jue Wang, Bhavana Konda, Jourdain Hayward, Christine M McLeod, Helen X Chen, Elad Sharon, Megan Othus, Christopher W Ryan, Melissa Plets, Charles D Blanke, and Razelle Kurzrock. A phase ii basket trial of dual anti–ctla–4 and anti–pd–1 blockade in rare tumors (dart) swog s1609: High-grade neuroendocrine neoplasm cohort. *Cancer*, 127(17):3194–3201, 2021.

Stephen Petersdorf, Kenneth Kopecky, Robert K Stuart, Richard A Larson, Thomas J Nevill, Leif Stenke, Marilyn L Slovak, Martin S Tallman, Cheryl L Willman, Harry Erba, and Frederick R Appelbaum. Preliminary results of southwest oncology group study s0106: An international intergroup phase 3 randomized trial comparing the addition of gemtuzumab ozogamicin to standard induction therapy versus standard induction therapy followed by a second randomization to post-consolidation gemtuzumab ozogamicin versus no additional therapy for previously untreated acute myeloid leukemia. *Blood*, 114(22):790, 2009.

Steven Piantadosi and Guanghan Liu. Improved designs for dose escalation studies using pharmacokinetic measurements. *Statistics in medicine*, 15(15):1605–1618, 1996.

Seth Pollack, Mary Weber Redman, Michael Wagner, Elizabeth Trice Loggers, Kelsey K Baker, Sabrina McDonnell, Jeffrey Gregory, Vanessa C Copeland, Kathryn J Hammer, Rylee Johnson, Roxanne Moore, Michael Shahnazari, Steven M Townson, Robin L Jones, and Lee D Cranmer. A phase i/ii study of pembrolizumab (pem) and doxorubicin (dox) in treating patients with metastatic/unresectable sarcoma., 2019.

John D Powderly, Bartosz Chmielowski, Julie R Brahmer, Sarina Anne Piha-Paul, Samantha Elizabeth Bowyer, Patricia LoRusso, Daniel VT Catenacci, Christina Wu, Minal A Barve, Michael Jon Chisamore, Nicole Nasrah, Dan Johnson, and William Ho. Phase i/ii dose-escalation and expansion study of flx475 alone and in combination with pembrolizumab in advanced cancer., 2020.

Matthew A Psioda, Jiawei Xu, QI Jiang, Chunlei Ke, Zhao Yang, and Joseph G Ibrahim. Bayesian adaptive basket trial design using model averaging. *Biostatistics*, 22(1):19–34, 2021.

Mark J Ratain. Redefining the primary objective of phase i oncology trials. *Nature Reviews Clinical Oncology*, 11(9):503–504, 2014.

Amanda J Redig and Pasi A Jänne. Basket trials and the evolution of clinical trial design in an era of genomic medicine. *J Clin Oncol*, 33(9):975–977, 2015.

Paul M Ridker, Mary Cushman, Meir J Stampfer, Russell P Tracy, and Charles H Hennekens. Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. *New England journal of medicine*, 336(14):973–979, 1997.

Caroline Robert, Jacob Schachter, Georgina V Long, Ana Arance, Jean Jacques Grob, Laurent Mortier, Adil Daud, Matteo S Carlino, Catriona McNeil, Michal Lotem, James Larkin, Paul Lorigan, Bart Neyns, Christian U Blank, Omid Hamid, Christine Mateus, Ronnie Shapira-Frommer, Michele Kosh, Honghong Zhou, Nageatte Ibrahim,

Scot Ebbinghaus, and Antoni Ribas. Pembrolizumab versus ipilimumab in advanced melanoma. *New England Journal of Medicine*, 372(26):2521–2532, 2015.

André Rogatko, James S Babb, Mourad Tighiouart, Fadlo R Khuri, and Gary Hudes. New paradigm in dose-finding trials: patient-specific dosing and beyond phase i. *Clinical cancer research*, 11(15):5342–5346, 2005.

Neil P Shah, Hagop M Kantarjian, Dong-Wook Kim, Delphine Réa, Pedro E Dorlhiac-Llacer, Jorge H Milone, Jorge Vela-Ojeda, Richard T Silver, H Jean Khoury, Aude Charbonnier, Nina Khoroshko, Ronald L Paquette, Michael Deininger, Robert H Collins, Irma Otero, Hughes Timothy, Eric Bleickardt, Lewis Strauss, Stephen Francis, and Andreas Hochhaus. Intermittent target inhibition with dasatinib 100 mg once daily preserves efficacy and improves tolerability in imatinib-resistant and-intolerant chronic-phase chronic myeloid leukemia. *Journal of Clinical Oncology*, 26(19):3204–3212, 2008.

Alexander D Sherry, Tharakeswara K Bathala, Suyu Liu, Bryan M Fellman, Stephen G Chun, Nikesh Jasani, B Ashleigh Guadagnolo, Anuja Jhingran, Jay P Reddy, Paul G Corn, Amishi Y Shah, Kelsey W Kaiser, Amol J Ghia, Daniel R Gomez, and Chad Tang. Definitive local consolidative therapy for oligometastatic solid tumors: Results from the lead-in phase of the randomized basket trial extend. *International Journal of Radiation Oncology* Biology* Physics*, 2022.

Jonathan Shuster. Optimal two-stage designs for single arm phase ii cancer trials. *Journal of Biopharmaceutical Statistics*, 12(1):39–51, 2002.

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

Richard Simon, Larry Rubinstein, Susan G Arbuck, Michaele C Christian, Boris Freidlin, and Jerry Collins. Accelerated titration designs for phase i clinical trials in oncology. *Journal of the National Cancer Institute*, 89(15):1138–1147, 1997.

Richard Simon, Susan Geyer, Jyothi Subramanian, and Sameek Roychowdhury. The bayesian basket design for genomic variant-driven phase ii trials. In *Seminars in Oncology*, volume 43, pages 13–18. Elsevier, 2016.

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.

David C Smith, Thomas Gajewski, Omid Hamid, Jeffrey S Wasser, Anthony J Olszanski,

Sandip P Patel, Ronac Mamtani, Emmett V Schmidt, Yufan Zhao, Janet E Maleski, and Tara C Gangadhar. Epacadostat plus pembrolizumab in patients with advanced urothelial carcinoma: Preliminary phase i/ii results of echo-202/keynote-037., 2017.

V Subbiah, RJ Kreitman, ZA Wainberg, JY Cho, JHM Schellens, JC Soria, PY Wen, CC Zielinski, ME Cabanillas, A Boran, P Ilankumaran, P Burgess, TR Salas, and B Keam. Dabrafenib plus trametinib in patients with braf v600e-mutant anaplastic thyroid cancer: updated analysis from the phase ii roar basket study. *Annals of Oncology*, 33(4):406–415, 2022.

Nicholas L Syn, Michele WL Teng, Tony SK Mok, and Ross A Soo. De-novo and acquired resistance to immune checkpoint targeting. *The Lancet Oncology*, 18(12):e731–e741, 2017.

Kentaro Takeda, Masataka Taguri, and Satoshi Morita. Boin-et: Bayesian optimal interval design for dose finding based on both efficacy and toxicity outcomes. *Pharmaceutical statistics*, 17(4):383–395, 2018.

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

Hussein Abdul-Hassan Tawbi, Weiyi Peng, Denai Milton, Rodabe Navroze Amaria, Isabella Claudia Glitza, Wen-Jen Hwu, Sapna Pradyuman Patel, Michael KK Wong, Scott Eric Woodman, Cassian Yee, Jennifer L McQuade, Michael T Tetzalaff, Alexander J Lazar, Suzanne Cian, Elizabeth M Burton, Jan H Beumer, Patrick Hwu, and Michael A Davies. Phase i/ii study of the pi3kβ inhibitor gsk2636771 in combination with pembrolizumab (p) in patients (pts) with pd-1 refractory metastatic melanoma (mm) and pten loss., 2018.

Matthew H Taylor, Chung-Han Lee, Vicky Makker, Drew Rasco, Corina E Dutcus, Jane Wu, Daniel E Stepan, Robert C Shumaker, and Robert J Motzer. Phase ib/ii trial of lenvatinib plus pembrolizumab in patients with advanced renal cell carcinoma, endometrial cancer, and other selected advanced solid tumors. *Journal of Clinical Oncology*, 38(11):1154, 2020.

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

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.

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

Peter F Thall, Richard M Simon, and Elihu H Estey. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in medicine*, 14 (4):357–379, 1995.

Peter F Thall, J Kyle Wathen, B Nebiyou Bekele, Richard E Champlin, Laurence H Baker, and Robert S Benjamin. Hierarchical bayesian approaches to phase ii trials in diseases with multiple subtypes. *Statistics in medicine*, 22(5):763–780, 2003.

Peter F Thall, Leiko H Wooten, Christopher J Logothetis, Randall E Millikan, and Nizar M Tannir. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in medicine*, 26(26): 4687–4702, 2007.

Peter F Thall, Hoang Q Nguyen, Thomas M Braun, and Muzaffar H Qazilbash. Using joint utilities of the times to response and toxicity to adaptively optimize schedule–dose regimes. *Biometrics*, 69(3):673–682, 2013.

D Thomas, D Chancellor, A Micklus, S LaFever, M Hay, S Chaudhuri, R Bowden, and AW Lo. Clinical development success rates and contributing factors 2011–2020. *Biotechnology Innovation Organization, Informa Pharma Intelligence, Quantitative Life Sciences, Washington, DC*, 2021.

Karen Tumaneng, Ryan C Russell, and Kun-Liang Guan. Organ size control by hippo and tor pathways. *Current Biology*, 22(9):R368–R379, 2012.

Moreno Ursino, Sarah Zohar, Frederike Lentz, Corinne Alberti, Tim Friede, Nigel Stallard, and Emmanuelle Comets. Dose-finding methods for phase i clinical trials using pharmacokinetics in small populations. *Biometrical Journal*, 59(4):804–825, 2017.

Aria Vaishnavi, Anh T Le, and Robert C Doebele. Trking down an old oncogene in a new era of targeted therapy. *Cancer discovery*, 5(1):25–34, 2015.

Paul A Volberding, Stephen W Lagakos, Matthew A Koch, Carla Pettinelli, Maureen W Myers, David K Booth, Henry H Balfour Jr, Richard C Reichman, John A Bartlett, Martin S Hirsch, Robert L Murphy, W.David Hardy, Ruy Soeiro, Margaret A Fischi, John G Bartlett, Thomas C Merigan, Hyslopm Newton E, Douglas D Richman, Fred T Valentine, and Lawrence Corey. Zidovudine in asymptomatic human immunodeficiency virus infection: a controlled trial in persons with fewer than 500 cd4-positive cells per cubic millimeter. *New England Journal of Medicine*, 322(14):941–949, 1990.

Nolan A Wages and Mark R Conaway. Phase i/ii adaptive design for drug combination oncology trials. *Statistics in medicine*, 33(12):1990–2003, 2014.

Jon Wakefield. The bayesian analysis of population pharmacokinetic models. *Journal of the American Statistical Association*, 91(433):62–75, 1996.

Ronald L Wasserstein and Nicole A Lazar. The asa statement on p-values: context, process, and purpose, 2016.

J Kyle Wathen, Peter F Thall, John D Cook, and Elihu H Estey. Accounting for patient heterogeneity in phase ii clinical trials. *Statistics in medicine*, 27(15):2802–2815, 2008.

Patrick Y Wen, Alexander Stein, Martin van den Bent, Jacques De Greve, Antje Wick, Filip YFL de Vos, Nikolas von Bubnoff, Myra E van Linde, Albert Lai, Gerald W Prager, Mario Campone, Angelica Fasolo, Jose A Lopez-Martin, Tae Min Kim, Warren P Mason, Ralf-Dieter Hofheinz, Jean-Yves Blay, Daniel C Cho, Anas Gazzah, Damien Pouessel, Jeffrey Yachnin, Aislyn Borach, Paul Burgess, Palanichamy Ilankumaran, Eduard Gasal, and Vivek Subbiah. Dabrafenib plus trametinib in patients with brafv600e-mutant low-grade and high-grade glioma (roar): A multicentre, open-label, single-arm, phase 2, basket trial. *The Lancet Oncology*, 23(1):53–64, 2022.

Howard Jack West. Novel precision medicine trial designs: umbrellas and baskets. *JAMA oncology*, 3(3):423–423, 2017.

John Whitehead, Yinghui Zhou, Lisa Hampson, Edouard Ledent, and Alvaro Pereira. A bayesian approach for dose-escalation in a phase i clinical trial incorporating pharmacodynamic endpoints. *Journal of Biopharmaceutical statistics*, 17(6):1117–1129, 2007.

Shinjo Yada and Chikuma Hamada. A bayesian hierarchal modeling approach to shortening phase i/ii trials of anticancer drug combinations. *Pharmaceutical statistics*, 17 (6):750–760, 2018.

F Yan, PF Thall, KH Lu, MR Gilbert, and Y Yuan. Phase i–ii clinical trial design: a state-of-the-art paradigm for dose finding. *Annals of Oncology*, 29(3):694–699, 2018.

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.

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.

Ying Yuan and Valen E Johnson. Bayesian hypothesis tests using nonparametric statistics. *Statistica Sinica*, pages 1185–1200, 2008.

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.

Ying Yuan and Guosheng Yin. Bayesian phase i/ii adaptively randomized oncology trials with combined drugs. *The annals of applied statistics*, 5(2A):924, 2011.

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 trialsboin: A simple, well-performing bayesian phase i design. *Clinical Cancer Research*, 22(17): 4291–4301, 2016.

Ying Yuan, Ruitao Lin, Daniel Li, Lei Nie, and Katherine E Warren. Time-to-event bayesian optimal interval design to accelerate phase i trials. *Clinical Cancer Research*, 24(20):4921–4930, 2018.

Yong Zang and J Jack Lee. A robust two-stage design identifying the optimal biological dose for phase i/ii clinical trials. *Statistics in medicine*, 36(1):27–42, 2017.

Jin Zhang and Thomas M Braun. A phase i bayesian adaptive design to simultaneously optimize dose and schedule assignments both between and within patients. *Journal of the American Statistical Association*, 108(503):892–901, 2013.

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.

Yanhong Zhou, J Jack Lee, and Ying Yuan. A utility-based bayesian optimal interval (u-boin) phase i/ii design to identify the optimal biological dose for targeted and immune therapies. *Statistics in medicine*, 38(28):S5299–S5316, 2019.

VITA

Mengyi Lu was born and grew up in Nanjing, Jiangsu, China. After graduating high school, she entered Arizona State University, Tempe, Arizona, USA. In 2015, she got her bachelor's degree in economics. In 2017, she entered Northeastern University to pursue her master's degree. In August 2019, she joined the University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences for a Ph.D. degree in biostatistics.