8-2017

NOVEL BAYESIAN ADAPTIVE CLINICAL TRIAL DESIGNS IN EARLY PHASES

Haitao Pan

NOVEL BAYESIAN ADAPTIVE CLINICAL TRIAL DESIGNS IN EARLY PHASES

by

Haitao Pan, Ph.D.

APPROVED:

Ying Yuan, Ph.D.

Advisory Professor

Jing Ning, Ph.D.

Xuelin Huang, Ph.D.

Yisheng Li, Ph.D.

Clifton David Fuller, M.D./Ph.D.

APPROVED:

Dean, The University of Texas

MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

NOVEL BAYESIAN ADAPTIVE CLINICAL TRIAL DESIGNS IN EARLY PHASES

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

Haitao Pan, Ph.D.

Houston, Texas, USA

August, 2017

DEDICATION

To my families

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisors, Dr. Ying Yuan for all his guidance and patience. He has been there for me every step of the way, encouraging and helping, as mentors and friends. I would also like to thank my committee members, Dr. Yisheng Li, Dr. Xuelin Huang, Dr. Jing Ning, and Dr. David Fulton Fuller, for their valuable comments and suggestions. Many professors and friends in M.D.Anderson Cancer Center also gave me tremendous help, my special thanks to Dr. Jack Lee, Dr. Suyu Liu, Dr. Ken Chen, Dr. Xiaoping Su, Dr. Shouhao Zhou, Dr. Nabihah Tayob, Dr. Sheng Luo, Dr. Hongjian Zhu, Dr. Hongyu Miao, Dr. Junsheng Ma, Mrs. Rong Ye, Mrs. Wei Qiao, Mrs. Hsiang-Chun Chen, Mrs. LeeAnn Chastain, and Mrs. Amy E. Carter.

I feel so lucky to be surrounded by many loving officemates, Heng Zhou, Youyi Zhang, Xiao Su, Xiao Li, Tianjiao Dai, Jing Piao, Xuebei An, Meiling Huang, Liangcai Zhang, Rongji Mu, Xuan Zhu and Jiong Chen etc., during the three years. All the happiness and laugh shared with them will stay in my memory forever.

My special thanks go to my wife, Chen, for his support over all these years.

ABSTRACT

NOVEL BAYESIAN ADAPTIVE CLINICAL TRIAL DESIGNS IN EARLY PHASES

Haitao Pan, Ph.D.

Advisory Professor: Ying Yuan, Ph.D.

Early phase, or phase I and phase II, trials are the first step in testing new medicines that have been developed in the lab. The main goal of phase I clinical trials is to establish the recommended dose of new drugs for phase II trials. For the cytotoxic drugs, the goal is to find maximum tolerated dose (MTD). The guiding principle for dose escalation in phase I trials is to avoid exposing too many patients to subtherapeutic doses while preserving safety and maintaining rapid accrual. Therefore, dose escalation methods, especially Bayesian designs, are recommended to be used in phase I trials. There are many proposed Bayesian phase I adaptive designs, among them, continual reassessment method (CRM) is the firstly proposed pioneered Bayesian design. The CRM needs pre-specification of a series of prior guesses of toxicity probabilities of each investigated doses, known as the skeleton, using a parametric model, and then continuously updates the estimate of the dose-toxicity curve based on accumulating data. By using a dose-toxicity model, the CRM efficiently pools data across doses and adaptively makes the decision of dose assignment and selection. Two chapters of the thesis devote to development of the CRM design (chapter 2) and to extend the CRM design (chapter 3). Specifically, chapter 2 deals with the issue of skeleton pre-specification in the CRM design. We propose an automatic

way to generate multiple skeletons for Bayesian model averaging CRM (BMA-CRM), an extension of robust version of the CRM, to avoid arbitrary specification of skeletons with improving performances compared to the original CRM and BMA-CRM designs. Chapter 3 deals with bridging studies, or follow-up trials. The emergence of bridging studies is due to different ethnic populations with different responses to a same drug and consequentially attaining different MTDs. Therefore, conventional one-size-fit-all paradigm cannot work. But, despite variations among different ethnic populations, their drug responses still show somewhat similarities. Commonly, a landmark trial has been conducted and a MTD dose has also been established for a certain population. Thus, independent conducting a trial for a new population of ignoring information of the landmark trial is wasteful. Therefore, challenges of the bridging studies include: how to effectively use/borrow information of the historical landmark trial, and how to design trials to accommodate heterogeneities of different populations. In this chapter, we develop a novel design, Bridging-CRM, B-CRM, to borrow the landmark trial information based on a proposed mixture estimator and the CRM framework, and to acknowledge different populations' heterogeneities of using the idea of multiple skeletons. Chapter 4 focuses on phase II design for biosimilar drug development. Biosimilar is a term that describes the equivalence of a generic version to an innovator's biologic drug product; biosimilars are close, but not exact copies of biologic drugs already on the market. Guidelines for statistical methods to establish biosimilarity remain nonspecific because of the newness of biosimilars. It is therefore of high urgency to develop appropriate and reliable statistical methodologies for developing biosimilars. Some statistical methods have been proposed to assess biosimilarity, but none of them proposed designs in this field. However, biosimilar trials come with several challenges that are beyond the scope of the conventional randomized comparative trial design. First, when a biosimilar is ready to be tested in a randomized trial, the innovative reference drug has been in the market for many years

and a huge amount of data on that drug has accumulated. It is critical to incorporate these rich historical data into the biosimilar trial design to improve trial efficiency. Another challenge when designing biosimilar trials is determining how to quantify and monitor the biosimilar during the trial. To address these issues, in chapter 4, we develop a new approach, the *calibrated power prior* (CPP), to allow the design to adaptively borrow information from the historical data according to the congruence between the historical data and the data collected in the current trial. We also propose the *Bayesian biosimilarity index* (BBI) to assess the similarity between the biosimilar and the innovative reference drug. In our design, we evaluate the BBI in a group sequential fashion based on the accumulating interim data, and stop the trial early once there is enough information to conclude or reject the similarity.

TABLE OF CONTENTS

Page

Page

List of Figures

List of Tables

1. Introduction

1.1 Background

The primary goal of a phase I clinical trial is to identify the maximum tolerated dose (MTD) of a new drug, which is defined as the dose with the toxicity probability closest to the target toxicity rate. Numerous statistical methods have been developed for phase I dose-finding studies, for example, the conventional 3+3 design [1], the continual reassessment method (CRM) [2], the decision-theoretic approach [3], the dose escalation method with overdose control [4], the improved up-and-down design [5], biased coin design [6], sequential testing approach [7], the modified toxicity probability interval design [8], and the newly developed Bayesian optimal interval design [9], among others.

The CRM is an important phase I trial design that pioneered the model-based adaptive dose-finding approach. The CRM prespecifies an initial shape of the dose-toxicity curve, known as the skeleton, using a parameter model, and then continuously updates the estimate of the dose-toxicity curve based on accumulating data. The CRM adaptively makes the decision of dose assignment and selection. By using a dose-toxicity model, the CRM efficiently pools data across doses; however, this means that it is also subject to the effects of model misspecification. Although the CRM is generally robust [10, 11], its performance can be undermined if the skeleton substantially deviates from the true dose-toxicity curve [12, 13, 14]. Before the phase I trial is conducted, there is typically limited prior information on the shape of the skeleton and thus the prespecification of the skeleton can be rather arbitrary. To overcome this issue, Yin and Yuan [12] proposed Bayesian model averaging CRM

(BMA-CRM), which specifies multiple skeletons, say three. By treating each skeleton as an independent model, the BMA-CRM uses a Bayesian model averaging or selection approach to automatically favor the best fitting skeleton and thus improve the robustness of the CRM. Standalone software with a graphic user interface for BMA-CRM is freely available at the MD Anderson Department of Biostatistics. The software has been downloaded more than 700 times since its completion. BMA-CRM has been used for a number of ongoing phase I trials at MD Anderson Cancer Center and other institutions.

The most common question we have received from the users of the BMA-CRM is how to specify the three skeletons. This is a daunting task for most practitioners. Specifying even a single skeleton for the standard CRM is not an easy task in practice. However, based on the desirable results of using multiple skeletons in the paper [12], it's worthwhile to refine the current version of the BNA-CMR method. Lee and Cheung [15] proposed a systematic method to simplify the specification of the skeleton of the CRM on the basis of the indifference interval. That method works well in practice, however, it cannot be directly used to specify multiple skeletons. The rationale of the BMA-CRM is to use different skeletons to cover different possible shapes of the dose-response curve, such that as long as one of them is close to the true dose-toxicity curve, the BAM-CRM will perform well. Therefore, the natural guidance is that we should choose the skeletons in such a way that each of them represents different shapes of the dose-toxicity curve [12]. For example, we can set a skeleton to represent a slowly increasing dose-toxicity curve with a high dose as the MTD; while another skeleton represents a quickly increasing dose-toxicity curve with a low dose as the MTD. However, as we show, the specification of multiple skeletons is actually more complicated than that because seemingly different skeletons can lead to an equivalent model.

1.2 Research One

In the chapter 2, we propose an automatic method to help practitioners specify multiple skeletons for BMA-CRM. We first define the equivalence of multiple skeletons and then convert the problem of measuring the equivalence of skeletons into a collinearity problem. Combining the proposed nonequivalence measure of multiple skeletons with the calibration method proposed by Lee and Cheung [15], we devise an automatic way to specify the optimal multiple skeletons that maximize the average percentage of correct selection of the MTD and meanwhile ensure sufficient nonequivalence among the skeletons. Simulation studies show that the proposed method has desirable operating characteristics. Software to implement the proposed method is available for free downloading at *www.trialdesign.org*.

1.3 Research Two

Traditionally, phase I trials are conducted in a "one-size-fits-all" fashion. That is, once the MTD of a drug is established in a landmark study based on a certain ethnic population (e.g., a Caucasian population), the results are directly extrapolated to other ethnic populations (e.g., an Asian population). Unfortunately, accumulating evidence shows that such a one-size-fits-all dose-finding paradigm is problematic, and ethnicity plays an important role in a patient's response to a drug [21]. The genetic and environmental differences among ethnic populations influence both the pharmacokinetics and pharmacodynamics of drugs [22]. As a result, different ethnic populations may have different MTDs. For example, a recent study of sorafenib administered in patients with hepatocellular carcinoma (HCC) has found that the MTD of sorafenib is significantly lower among Asian patients than among non-Asian patients [23]. According to the manufacturer and the United States Food and Drug Administration (FDA), the recommended dose of sorafenib is 400 mg bid. That dose

has been used in the pivotal phase III Sorafenib HCC Assessment Randomized Protocol (SHARP) trial, which involves a patient population drawn from Europe, North America and South America [24]. However, the study of Barrera et al. [23] showed that Asian patients demonstrated poor tolerance to the manufacturer's recommended initial dose of the drug. Among a total of 36 Asian patients evaluated with this drug, 97% did not tolerate the FDA-indicated dose of 400 mg bid. Another example of inter-ethnic difference in drug tolerance is the administration of docetaxel, a broad spectrum taxane commonly used to treat solid tumors, such as lung, breast, gastric, ovarian and prostate tumors. The main side effects of docetaxel are myelosuppression and peripheral neuropathy. Different doses of docetaxel are administered in different geographic populations. In Caucasians, the common starting dose of docetaxel for the first-line treatment is 100mg/m^2 [25]; whereas in countries with Asian populations, such as China, the common starting dose is 70 to 75mg/m^2 [26], and in Japan, the approved starting dose for docetaxel is 60 mg/m^2 [27]. Despite using the reduced doses, the incidence of febrile neutropenia was still higher among Asians than among Caucasians [26, 27]. Such inter-ethnic differences in docetaxel tolerance may arise from different clearance and exposure rates in Asians and Caucasians [28]. Goh et al. [29] reported that docetaxel clearance is approximately 40% lower, while the area under the curve for docetaxel (i.e., drug exposure) is approximately 25% higher in Asians than in Caucasians.

The inter-ethnic differences have being recognized by drug regulatory authorities. In 1999, the FDA published the guidance, "Ethnic Factors in the Acceptability of Foreign Clinical Data" (i.e., E5 Guidance), which suggested the need to distinguish three ethnic categories (i.e., Asian, Black, and Caucasian) in drug development. The guidance identified situations for which drugs could be ethnically sensitive and suggested the types of bridge studies that may be required to extrapolate clinical trial results from one region to another. In 2005, the FDA updated the guidance and extended

the race category further to include 5 minimum ethnic groups, namely, Caucasian, Black/African American, Asian, American Indian/Alaskan Native, and Native Hawaiian/other Pacific Islander. In the same year, the FDA approved the first race-specific drug, isosorbide dinitrate and hydralazine hydrochloride (proprietary name: BiDil), for the treatment of congestive cardiac failure in black patients.

The goal of the chapter 3 is to address the following bridge trial design question: given that a landmark phase I trial has been conducted in a landmark population and the corresponding MTD has been established, how do we design a follow-up trial (i.e., a bridge trial) to find the MTD for a new population. A straightforward approach is to ignore the early landmark trial and conduct another independent phase I trial (e.g., using the CRM) to find the MTD for the new population. While this approach fully acknowledges the inter-ethnic heterogeneity, it is not efficient because the dose-toxicity relationships in different ethnic populations are expected to be closely related, even though there are some inter-ethnic differences. This is true because, after all, we are concerned with the same drug acting through the same biological mechanism in human beings. In other words, the dose-toxicity response observed in the landmark trial should inform the basic dose-toxicity behavior of the drug in the new population. Ignoring the data from the landmark trial is wasting useful information.

To address this issue, we propose the bridging CRM (B-CRM) design, which utilizes the dose-toxicity data obtained from the landmark trial to achieve efficient dose finding in the follow-up trial, while also acknowledging inter-ethnic heterogeneity. Specifically, we first estimate the dose-toxicity curve using the data from the landmark trial, and then use that estimate to form a prior dose-toxicity curve, which is also known as the skeleton of the CRM, for the follow-up trial. To accommodate the inter-ethnic heterogeneity, we form multiple skeletons, by shifting the estimate of the dose-toxicity curve one dose level up or down, to represent a more conservative or aggressive dose response in the new population. We employ the Bayesian model

averaging approach [**?**] to draw inference across multiple skeletons and adaptively make the decision of dose assignment and selection. This article focuses on ethnic heterogeneity, but the proposed method can be used to handle other types of patient heterogeneity, e.g., patient subgroups defined by prognostic factors or biomarkers.

Some research has been done for bridging studies. Shih [30] proposed a method to determine whether a study is capable of bridging the foreign data to the new study. Lan et al. [31] proposed weighted Z-tests in which incorporate the prior observed data in bridging studies using weights. Gould et al. [32] developed a Bayesian predictive approach for designing and analyzing bridging studies that fully incorporates the information provided by the original trials. Gandhi et al. [33] proposed a Bayesian approach for inference from a bridging study with binary outcomes. Chow et al. [34] provided a review of statistical methods for bridging studies. Most of the aforementioned works focus on statistical inference, and limited research has been done from the trial design perspective. Morita [35] proposed a phase I trial design that uses an informative prior to incorporate previous study information into the bridging study based on the CRM.

1.4 Research Three

According to the Patient Protection and Affordable Care Act (Affordable Care Act) [47], signed into law by President Obama, a biosimilar product is defined as "a biological product that is highly similar to its reference product notwithstanding minor differences in clinically inactive components and there are no clinically meaningful differences in terms of safety, purity, and potency." In other words, biosimilar is a term that describes the equivalence of a generic version to an innovator's biologic drug product; biosimilars are close, but not exact copies of biologic drugs already on the market. Examples of biological products include vaccines, blood products for transfusion, human cells and tissues used for transplantation, gene therapies, and cel-

lular therapies [48]. Many biological drugs are important life-saving products but are extremely expensive, which severely limits their accessibility to the general patient population. As the patents of many blockbuster proprietary biologic products reach their expirations, such as those for rituximab, infliximab, palivizumab, omalizumab and trastuzumab, biosimilars provide great potential to increase the accessibility of biologic products for patients with life-threatening diseases. Currently, more than 80 biosimilars are under development, and global sales of biosimilars have been estimated to reach $3.7 billion in 2015 [49].

Before a biosimilar can be used to treat patients, it must demonstrate "biosimilarity" to its innovative reference product in terms of quality characteristics, biological activity, safety and efficacy based on comprehensive comparability studies [50]. Because the development of biological products is much more complicated than that of conventional small-molecule-based drugs, and biologics are sensitive to small procedural or environmental changes during the manufacturing process, the conventional approach to evaluating bioequivalence based on pharmacokinetic responses cannot be directly applied to establish biosimilarity. Biosimilars cannot be regarded as generic equivalents (or biogenerics) of innovative drugs because of the impossibility of the active ingredients in biosimilars being identical to their innovative counterparts [51]; whereas generic small-molecule drugs can be considered therapeutically equivalent to an innovative drug if pharmaceutical equivalence and bioequivalence can be demonstrated. Guidelines for statistical methods to establish biosimilarity remain nonspecific because of the newness of biosimilars, even though regulatory agencies, such as the U.S. Food and Drug Administration (FDA), the European Medicines Agency, and the World Health Organization, have provided detailed guidance for demonstrating comparability in terms of quality, safety and efficacy. It is therefore of high urgency to develop appropriate and reliable statistical methodologies for developing biosimilars.

Some statistical methods have been proposed to assess biosimilarity. Lin et al. [52] proposed a way to assess biosimilar products for binary endpoints using a parallel line assay method; Li et al. [61] proposed a method for considering biosimilar clinical efficacy trials with asymmetrical margins; Kang et al. [60] proposed a similarity criterion using a relative distance method based on the absolute mean difference between a biosimilar product and the innovative reference product; Chow et al. [57,58,59] made important comments and discussed several scientific and practical issues raised in the FDA guidance; Endrenyi et al. [56] discussed the differences between small-molecule drugs and biologicals with respect to the interchangeability of drug products; for the quality control of biosimilars, Yang et al. [55] proposed an adapted F-test for homogeneity of the variances to assess biosimilarity in variability; and that issue was also considered by Zhang et al. [54] and Liao et al. [53]. Combest et al. [63] reviewed the existing methods and demonstrated on a conceptual level that a Bayesian approach can reduce the sample size compared to the traditional frequentist approach and batch-to-batch methods when developing a biosimilar. These existing methods have mainly focused on the statistical assessment of biosimilarity; little research has been done on designing clinical trials for biosimilars, especially from the Bayesian perspective. A monograph by Chow [62] provides an excellent review of biosimilar drug development.

In the chapter 4, we propose a two-arm randomized Bayesian group sequential design to evaluate the biosimilarity between an investigational biosimilar and the innovative reference drug. Biosimilar trials come with several challenges that are beyond the scope of the conventional randomized comparative trial design. First, when a biosimilar is ready to be tested in a randomized trial, the innovative reference drug has been in the market for many years and a huge amount of data on that drug has accumulated. It is critical to incorporate these rich historical data into the biosimilar trial design to improve trial efficiency. An efficient trial design not only leads to

tremendous cost saving for the pharmaceutical industry, but translates into saving lives because it allows patients to access the biosimilars earlier by expediting their development. Another challenge when designing biosimilar trials is determining how to quantify and monitor the biosimilar during the trial. To address these issues, we have developed a new approach, the *calibrated power prior* (CPP), to allow the design to adaptively borrow information from the historical data according to the congruence between the historical data and the data collected in the current trial. We have also proposed the *Bayesian biosimilarity index* (BBI) to assess the similarity between the biosimilar and the innovative reference drug. In our design, we evaluate the BBI in a group sequential fashion based on the accumulating interim data, and stop the trial early once there is enough information to conclude or reject the similarity. Simulation studies show that our method is statistically powerful, with well controlled type I error rates.

This thesis interpolates material from three papers by the author [80,81,82]. Chapter 2 uses material from Reference [80], coauthored with Ying Yuan. Meanwhile, Chapter 3 is based on Reference [81], coauthored with Suyu Liu, Jielai Xia, Qin Huang and Ying Yuan. Finally, Chapter 4 is based on Reference [82], coauthored with Ying Yuan and Jielai Xia. Some material from each of these papers has also been incorporated into this introductory Chapter.

2. A default method to specify skeletons for Bayesian model averaging continual reassessment method for phase I clinical trials

2.1 Introduction

This chapter is based on "A default method to specify skeletons for Bayesian model averaging continual reassessment method for phase I clinical trials" published in Statistics in Medicine (2016) [80] coauthored with Ying Yuan. Permission from the journal has been granted for use in conjunction with the thesis.

As introduced above, the primary goal of a phase I clinical trial is to identify the maximum tolerated dose (MTD) of a new drug and the CRM is an important phase I trial design that pioneered the model-based adaptive dose-finding approach. The BMA-CRM is an extension of the original version by specifying multiple skeletons to produce more robust results than the original CRM design. But, how to specify multiple skeletons is a challenging task.

In this chapter, we propose an automatic method to help practitioners specify multiple skeletons for BMA-CRM. We first define the equivalence of multiple skeletons and then convert the problem of measuring the equivalence of skeletons into a collinearity problem. Combining the proposed nonequivalence measure of multiple skeletons with the calibration method proposed by Lee and Cheung [15], we devise an automatic way to specify the optimal multiple skeletons that maximize the average percentage of correct selection of the MTD and meanwhile ensure sufficient nonequivalence among the skeletons. Simulation studies show that the proposed method has

desirable operating characteristics. Software to implement the proposed method is available for free downloading at *www.trialdesigns.org*.

The remainder of the chapter is organized as follows. In Section 2.2, after briefly reviewing the BMA-CRM and the calibration method of Lee and Cheung, we introduce the concept of equivalency of skeletons and the procedure to choose the optimal skeletons for the BMA-CRM. In Section 2.3, we investigate the operating characteristics of the proposed approach and conclude with a brief discussion in Section 2.4.

2.2 Method

2.2.1 Bayesian model averaging CRM (BMA-CRM)

Let $d_1 < \cdots < d_J$ denote a set of J prespecified doses of a new drug under investigation, and ϕ be the target toxicity rate. The standard CRM assumes a working dose-toxicity model, and then based on the accumulating data, continuously updates the estimate of the dose-toxicity model and makes the decision of dose escalation/de-escalation. A commonly used working model is the following power (or empirical) model,

$$\Pr(\text{toxicity at } d_j) = \pi_j = p_j^{\exp(\alpha)}, \qquad j = 1, \cdots, J \tag{2.1}$$

where α is an unknown parameter, and $(p_1, \cdots, p_J)$ are a set of prespecified constants, known as skeletons. The skeleton can be interpreted as the prior guess of toxicity probabilities at J doses. A normal prior distribution $N(0, \sigma^2)$ is often assumed for α, e.g., $\alpha \sim N(0, 2)$.

Although the CRM is generally robust, it is still influenced by the misspecification of the dose-toxicity model. If the assumed dose-toxicity model substantially deviates from the true dose-toxicity relationship, the performance of the CRM can be

compromised under small sample sizes. The BMA-CRM addresses this issue by specifying multiple skeletons, each of them leading to a dose-toxicity model, and then using Bayesian model averaging to automatically favor the best fitting model for the decision of dose escalation/de-escalation. Specifically, let $\{(p_{11}, \cdots, p_{1J}), \cdots, (p_{K1}, \cdots, p_{KJ})\}$ denote K prespecified skeletons, and $(M_1, \cdots, M_K)$ be the corresponding models generated by each of these skeletons, with M_k given by

$$\pi_{kj} = p_{kj}^{\exp(\alpha_k)}. \tag{2.2}$$

Assume that at a certain stage of the trial, n_j patients are treated at dose j, and y_j of them experience dose-limiting toxicity (DLT), for $j = 1, \cdots, J$. Given the observed data $D = \{(n_j, y_j), j = 1, \cdots, J\}$, the likelihood function under model M_k is

$$L(D|\alpha_k, M_k) \propto \prod_{j=1}^{J} \{p_{kj}^{\exp(\alpha_k)}\}^{y_j} \{1 - p_{kj}^{\exp(\alpha_k)}\}^{n_j - y_j}. \tag{2.3}$$

Let $\Pr(M_k)$ be the prior probability that model M_k is the true model. The posterior model probability for M_k is given by

$$Pr(M_k|D) = \frac{L(D|M_k)Pr(M_k)}{\sum_{i=1}^{K} L(D|M_i)Pr(M_i)}, \tag{2.4}$$

where $L(D|M_k)$ is the marginal likelihood for model M_k, and

$$L(D|M_k) = \int L(D|\alpha_k, M_k) f(\alpha_k|M_k) d\alpha_k, \tag{2.5}$$

with α_k denoting the parameter of model M_k, and $f(\alpha_k|M_k)$ denoting the prior distribution of α_k under model M_k. When there is no preference *a priori* for any single model, we set equal prior probabilities $\Pr(M_k) = \frac{1}{K}$.

The BMA estimate for the toxicity probability at each dose level is given by

$$\bar{\pi}_j = \sum_{k=1}^{K} Pr(M_k|D)\hat{\pi}_{kj} \ , j = 1, \cdots, J, \tag{2.6}$$

where $\hat{\pi}_{kj}$ is the posterior mean of the toxicity probability of dose level j under model M_k, i.e.,

$$\hat{\pi}_{kj} = \int p_{kj}^{\exp(\alpha_k)} \frac{L(D|\alpha_k, M_k)f(\alpha_k|M_k)}{\int L(D|\alpha_k, M_k)f(\alpha_k|M_k)d\alpha_k} d\alpha_k. \tag{2.7}$$

By assigning $\hat{\pi}_{kj}$ a weight of $Pr(M_k|D)$, the BMA method automatically identifies and favors the best fitting model, thus $\bar{\pi}_j$ is always the best estimate. Based on $\bar{\pi}_j$, we can make the decision of dose escalation and de-escalation. The dose-finding algorithm for the BMA-CRM can be described as follows:

1. Patients in the first cohort are treated at the lowest dose d_1, or the physician-specified dose.

2. At the current dose level j^{curr}, we obtain the BMA estimates for the toxicity probabilities, $\bar{\pi}_j$ $(j = 1, \ldots, J)$, based on the cumulated data. We then find dose level j^* that has a toxicity probability closest to ϕ, i.e.,

$$j^* = \text{argmin}_{j\in(1,\ldots,J)}|\bar{\pi}_j - \phi|.$$

   If $j^{\text{curr}} > j^*$, we de-escalate the dose level to $j^{\text{curr}} - 1$; if $j^{\text{curr}} < j^*$, we escalate the dose level to $j^{\text{curr}} + 1$; otherwise, the dose stays at the same level as j^{curr} for the next cohort of patients.

3. Once the maximum sample size is reached, the dose that has the toxicity probability closest to ϕ is selected as the MTD.

In addition, we add a stopping rule in our algorithm:

$$\text{if pr(toxicity rate at } d_1 > \phi) = \sum_{k=1}^{K} \text{pr}\{\pi_{k1}(\alpha_k) > \phi | M_k, D\}\text{pr}(M_k|D) > 0.9,$$

the trial is terminated for safety.

2.2.2 Lee and Cheung's method for choosing a single skeleton

Lee & Cheung [15] proposed a practical method for choosing a single skeleton for the standard CRM based on the indifference interval, which is defined as an interval of toxicity probabilities associated with the neighboring doses of the true MTD such that these neighboring doses may be selected instead of the true MTD under large samples. In that approach, one specify the target probability of toxicity ϕ, the location of the MTD and an acceptable indifference interval δ. Then a skeleton will be uniquely determined and the chosen skeleton guarantees that the target probability of the DLT will fall in the specified indifference interval. This process can be conveniently implemented using function *getprior*$(\cdot)$ in R package "dfcrm". As the indifference interval is a large sample property, in order to ensure good performance in finite samples, Lee and Cheung [15] suggested numerically searching a range of acceptable indifference intervals, rather than prespecifying a fixed value of the indifference interval, to identify the optimal skeleton that yields the highest percentage of correct dose selection (PCS) based on a set of prespecified toxicity scenarios. The authors showed that this calibration method yields a skeleton with good operating characteristics. The method of Lee and Cheung is useful for selecting a single skeleton for the standard CRM, but cannot be used for selecting multiple skeletons. We will extend this method to selecting multiple skeletons, for example, which can be use for the BMA-CRM method.

2.2.3 Equivalency of skeletons

As introduced previously, the rationale behind the BMA-CRM is to use multiple skeletons (or models) to represent different dose-toxicity relationships, such that as long as one of them is close to the truth, we will obtain good design performance, thanks to the property that the BMA automatically identifies and favors the best fitting model. Therefore, ideally, we would like these skeletons to be as different as possible to maximize the coverage of the model space (i.e., all possible shapes of the dose-response relationship). Achieving this goal, however, is trickier than it appears. For example, Figure 2.1 shows three skeletons, which represent rather different prior opinions on the dose-toxicity profile. Skeleton 1 represents an aggressive prior opinion that the first dose is the MTD (with target toxicity probability of 0.3) and the dose-toxicity curve takes a concave shape; skeleton 2 represents a neutral prior opinion that the middle (i.e., 3rd) dose level is the MTD and toxicity increases linearly with the dose; skeleton 3 represents a conservative opinion that the dose starts with a low toxicity and the highest dose is the MTD. Although the three skeletons appear to be rather different, they are actually equivalent (see below). We use the following definition to determine equivalency of multiple skeletons.

Definition 1 Two skeletons $\boldsymbol{p} = \{p_1, \ldots, p_J\}$ and $\boldsymbol{p'} = \{p'_1, \ldots, p'_J\}$ are equivalent if $p'_j = p_j^c$, for $j = 1, \cdots, J$, where c is a constant.

This definition matters since the equivalent skeletons lead to equivalent dose-toxicity models (or likelihood) and thus the same posterior of estimate $\hat{\pi}_{kj}$. To see this, the model using skeleton $\boldsymbol{p'}$ is

$$\pi_j = (p_j^{'})^{\exp(\alpha)}.$$

Plugging in $p_j^{'} = p_j^c$, we obtain

$$\pi_j = {p_j}^{\exp(\alpha+\log(c))}. \tag{2.8}$$

Applying the reparameterization $\gamma = \alpha + \log(c)$, model (2.8) becomes

$$\pi_j = {p_j}^{\exp(\gamma)},$$

which is the same as the model that uses skeleton $\boldsymbol{p}$.

Letting $\boldsymbol{p} \sim \boldsymbol{p}'$ denote that $\boldsymbol{p}$ is equivalent to $\boldsymbol{p}'$, it is easy to see that equivalence has property of transitivity.

Lemma 1 If $\boldsymbol{p} \sim \boldsymbol{p}'$ and $\boldsymbol{p}' \sim \boldsymbol{p}''$, then $\boldsymbol{p} \sim \boldsymbol{p}''$.

The implication of the above results is that when the skeletons of the BMA-CRM are equivalent, it is the same as using a single skeleton and thus Bayesian model averaging cannot function for that purpose, i.e., to decrease the dependence of the CRM on a single skeleton and improve the robustness of the design. Hereafter, we discuss a way to measure the degree of equivalence among multiple skeletons and propose a method to optimize the choice of multiple skeletons.

An interesting application of the above Definition 1 and Lemma 1 is that we can prove that the skeletons generated by varying the MTD locations of parameter *nu* in the function *getprior* in the R package *dfcrm* proposed by Lee&Cheung (2009) are equivalent following the Definition 1.

Lemma 2 If $\boldsymbol{p}$ and $\boldsymbol{p}'$ are two skeletons generated by the method of Lee&Cheung. For these two skeletons, only initial guesses of location for the MTD are different, then $\boldsymbol{p} \sim \boldsymbol{p}'$.

A proof is in the Appendix of this chapter.

There are two points should be noted here. One is that in the paper of Jia et al. (2014), the authors discussed a concept "irrelevant", which is easy to be confused with our above proposed concept "invariant". In Theorem 1 of their paper, they proved that the choice of prior location v is "*irrelevant* to the performance of the likelihood continual reassessment method". Therefore, the "irrelevant" in Jia et al. paper, they proved that the performance of using the likelihood approach of the CRM design will not be influenced by the choose of parameter v, simply speaking, the "irrelevant" concept is associated with the final performance of the CRM. However, our "invariant" means that the ratios of logarithm of multiple skeletons are constant based on the different choice of parameter v. In this paper, the "invariant" is based on our proposed skeletons equivalence definition, which is totally different from Jia et al. "irrelevant" concept.

The other one point we should mention is that Jia et al. (2014 [11]) discussed many characterization of the CRM design. In their paper, they proposed a concept of $\psi-$equivalent functions. By using this concept, one can systematically expand the scope of the dose-toxicity functions for the CRM design instead of just focusing on the three commonly mentioned models of power function, hyperbolic tangent function and logistic function. Their work makes big theoretical contributions to the CRM design, however, our proposed "equivalence" aims for the selection of skeletons when using the CRM instead of choosing toxicity response models. Therefore, the two concepts are totally different. Readers should not be confused by them.

2.2.4 Optimize the choice of skeletons

As shown in Figure 2.1, though we now have the Definition 1, it is still difficult to evaluate the equivalency of multiple skeletons on the basis of a visual inspection. A quantitative measure that gauges the degree of equivalence among multiple skele-

ton needs to be developed. To do that, the key observation is that the equivalence condition $p'_j = p^c_j$ can be rewritten as

$$\log(p'_j) = c\log(p_j), \quad j = 1, \cdots, J.$$

Hence, if we view $\{\log(p'_j)\}$ and $\{\log(p_j)\}$ as observations from two independent random variables, the equivalency of two skeletons is the same as the logarithm of these two skeletons being in perfect collinearity. This result is simple but powerful because it converts the problem of determining the equivalence of skeletons into a well-studied, classical collinearity problem in linear regression analysis. Specifically, the problem of measuring the equivalence among K skeletons, $\boldsymbol{p}_1 = (p_{11}, \cdots, p_{1J})$, $\cdots$, $\boldsymbol{p}_K = (p_{K1}, \cdots, p_{KJ})$, can be converted into the problem of measuring the collinearity among K vectors $\boldsymbol{q}_1 = (\log(p_{11}), \cdots, \log(p_{1J}))$, $\cdots$, $\boldsymbol{q}_K = (\log(p_{K1}), \cdots, \log(p_{KJ}))$. Following Weinberg [17] (page 214-216), a common way to measure the collinearity among K skeletons, $\boldsymbol{q}_1, \cdots, \boldsymbol{q}_K$, is given by

$$\bar{R^2} = \frac{1}{K}\sum_{k=1}^{K} R_k^2, \tag{2.9}$$

where R_k^2 is the R^2 obtained by regressing $\boldsymbol{q}_k$ on the other $K-1$ skeletons, i.e., $\{\boldsymbol{q}_{k'}, k' \neq k\}$. The value of $\bar{R^2}$ is between 0 and 1. A small value indicates less collinearity, i.e., the K skeletons are less equivalent, and $\bar{R^2} = 1$ indicates perfect collinearity, i.e., the K skeletons are equivalent. Given a set of K skeletons, we define a measure for quantifying nonequivalency of the skeletons, denoted as Q, as follows,

$$Q = 1 - \bar{R^2} = \frac{1}{K}\sum_{k=1}^{K} R_k^2$$

.

Our method of choosing K skeletons for BMA-CRM is based on the nonequivalence measure Q and the skeleton calibration method of Lee and Cheung [15], which can be described as follows.

1. For each of K skeletons to be specified, generate a pool of candidate skeletons using Lee and Cheung's method based on a sequence of indifference intervals ranging from $[a, b]$ with a step of c. This results in K skeleton pools, each of which contains $S = (a - b)/c$ candidate skeletons.

2. Randomly select one skeleton from each of the K skeleton pools to form a K-skeleton set. This results in a total of S^K possible K-skeleton sets.

3. Calculate the value of the nonequivalence measure Q for each of the K-skeleton sets and sort them by the value of Q from large to small.

4. Pick the top 20 K-skeleton sets with the largest values of Q, and simulate 1,000 trials with each of them using the BMA-CRM under a set of prespecified toxicity scenarios. We choose the K-skeleton set that maximizes the average PCS (across the scenarios) as the recommended skeletons to be used in the BMA-CRM for conducting the actual trial.

Several remarks are warranted. In the above algorithm, we do not directly choose the K-skeleton set with the largest value of Q as the final recommended skeletons because, as pointed out by Lee and Cheung [15], the skeleton generated by the indifference interval only guarantees a good performance in a large sample. To ensure good finite-sample performance, following Lee and Cheung's approach, we choose the recommended skeleton as the skeleton set that yields the highest average PCS from the top 20 skeleton sets (i.e., step 4 of the algorithm). Note that, rather than maximizing the average PCS, other criteria, e.g., maximizing the lowest PCS among the simulation scenarios (i.e., the minimax criterion), can also be used to select the final recommended skeletons. The numerical studies we present show that the use of the

highest average PCS generally performs better than the minimax criterion. Last, in step 1, to generate a candidate skeleton, Lee and Cheung's method (i.e., function *getprior*$(\cdot)$ in R package "dfcrm") requires specifying three parameters: the target toxicity probability ϕ, the location of the MTD ν and an indifference interval δ. We know that ϕ have set the indifference interval as a sequence from $[a, b]$ with a step of c. To specify the location of the MTD (i.e., ν), we have proven in Lemma 2 that the skeleton generated by *getprior*$(\cdot)$ is actually equivalent under different values of ν. In other words, the skeleton generated by the method of Lee and Cheung is invariant to the location of the MTD. Therefore, without loss of generality, we simply set $\nu = [(k/K)J]$ (i.e., the MTD is the k/K percentile of the investigational doses) when generating the candidate skeleton pool for the kth skeleton, $k = 1, \cdots, K$.

2.3 Simulation studies

2.3.1 Operating characteristics

We used the simulation setting previously used by Yin and Yuan [12]. We assumed $J = 8$ dose levels, the target toxicity rate $\phi = 0.3$ and 8 toxicity scenarios (see Table 1). Following the recommendation of Yin and Yuan [12], we used 3 skeletons to run the BMA-CRM. To apply the proposed procedure to determine 3 skeletons, we set the indifference interval range as [0.02, 0.15], with a step of 0.01, i.e., $a = 0.02, b = 0.15$, and $c = 0.01$ in step 1 of the algorithm. This generated 14 different indifference intervals. Cheung [16] recommended the indifference interval range [0.04, 0.10] for the target toxicity rate of 0.33. We slightly expanded that range to [0.02, 0.15] to obtain a broader coverage of the skeleton space. We considered two versions of the proposed procedure, one maximizing the average PCS and one maximizing the lowest PCS. We refer to the two sets of resulting skeletons as optimal skeletons and minimax skeletons, respectively. We compared the performance of these two sets of skeletons

with a set of "empirical skeletons" specified by varying the prior location of the MTD and shape of the dose curve (see Figure 2.2):

Skeleton 1: 0.30, 0.39, 0.48, 0.57, 0.64, 0.71, 0.76, 0.81

Skeleton 2: 0.15, 0.19, 0.22, 0.26, 0.30, 0.34, 0.38, 0.42

Skeleton 3: 0.0001, 0.002, 0.01, 0.038, 0.095, 0.19, 0.30, 0.42

Although they appear to be very different, these three arbitrarily specified skeletons are actually close to being equivalent, with $Q = 0.03$.

Table 2.1 shows the results based on 1,000 simulated trials, including the selection percentage of each dose as the MTD, the average number of patients treated at each dose, and average number of toxicity events. In scenario 1, the third dose is the MTD. The empirical skeletons yielded the lowest PCS of 66.5% and the optimal skeletons yielded the highest PCS of 72.1%. The PCS achieved by the minimax skeletons is slightly lower than that achieved by the optimal skeletons (i.e., 70.3%). The number of patients treated at each dose using the empirical skeletons (12.8) is also less than the number obtained when using the optimal (13.2) and minimax (13.5) skeletons. Scenario 2 has the MTD at the fifth dose level, and the optimal skeletons performed best with the highest PCS (71.7%) and the largest number of patients treated at the MTD (10.1). The minimax skeletons performed the second best while the empirical skeletons performed the worst. Similar results are observed in scenarios 3, 4 and 5. In scenario 6, the sixth dose is the MTD, and the PCS obtained when using the optimal skeletons is the highest (40.3%), but that obtained when using the minimax skeletons is lowest (37%). Similar results are observed in scenario 7. In scenario 8, the MTD is located at the sixth dose. In this case, the empirical skeletons yielded the highest PCS of 45.3%. The PCS obtained when using the optimal skeletons is 43.3%, which is very close to the best one. The PCS obtained when using the minimax skeletons is

37.0%. The average numbers of patients treated at the MTD are similar among the three sets of skeletons.

For thorough exploration of our proposed method, we also compare our method to the CRM method using just one single skeleton. For the one single skeleton CRM, we use the *getprior()* to produce the skeletons. Specifically, we set the indifference to be 0.10, and set the location parameter $v = 5$ (if we set $v = 4$, the results are similar by our simulations). From Table 1, we can see that, if the dose level is evenly distributed, for example, in Scenarios 1,4,5, the CRM with one single skeleton have an average 9% higher PCS than those of our proposed method; however, in other scenarios (2,3,6,7,8), our proposed method produced much better performances, an average 128% higher PCS, than the CRM with just a single skeleton.

In summary, the simulations show that using the proposed optimal skeletons can improve the performance of the BMA-CRM. The minimax skeletons performed reasonably well but not as well as the optimal skeletons. The BMA-CRM equiping with our proposed skeletons choice procedure has better performance than using one single skeleton CRM design. We recommend the optimal skeletons of using the BMA-CRM for practical use.

2.3.2 Sensitivity analysis

Our algorithm picks the skeleton set that maximizes the average PCS from the top 20 skeleton sets (with the largest values of Q) as the recommended optimal skeletons. To investigate the effect of the nonequivalence measure Q on the performance of the design, we considered two other ways to choose the optimal skeletons. Specifically, after calculating the value of the nonequivalence measure Q for all possible K-skeleton sets and sorting them by the value of Q from large to small (i.e., step 3 of the algorithm), rather than picking the top 20 skeleton sets, we picked the middle or bottom 20 skeleton sets, among which we chose the skeleton set that maximized the

average PCS as the optimal skeletons. We compared the performance of these three ways of choosing the optimal skeletons under 8 scenarios. Figures 2.3 and 2.4 display the PCS and the number of patients treated at the MTD. We can see that in general, the optimal skeleton set selected from the sets with the top 20 Q values performed better than that based on the sets with the middle 20 Q values, which performed better than that based on the sets with the bottom 20 Q values. This result shows that using skeletons with larger values of Q, i.e., using more diverse skeletons, generally improves the performance of the BMA-CRM. Across the 8 scenarios, although the optimal skeleton set based on the sets with the top 20 Q values has the best or near best performance in terms of both the PCS and the number of patients treated at the MTD, it is not always the best. This is reasonable because, in the BMA-CRM, the objective of using multiple skeletons is to improve the robustness of the design and ensure that the design generally has good performance across various scenarios, not to guarantee that the design always perform the best in every single scenario. The details of the simulation results are provided in Table 2.

2.4 Summary

The BMA-CRM is an extension of the CRM that improves the robustness of the design by specifying multiple skeletons and then using Bayesian model averaging to automatically favor the best fitting model for robust dose finding. The major difficulty for practitioners when using the BMA-CRM is the requirement of specifying multiple skeletons. To overcome this issue, we propose a default, automatic method to help practitioner specify multiple skeletons when using the BMA-CRM. We define a measure to gauge the difference among multiple skeletons and then, based on that measure, we develop a model calibration method to select the optimal skeletons. The simulation studies show that the proposed method produces robust operating characteristics. The algorithm is straightforward and easy to implement, and we

provide the R function to facilitate the use of the method in practice. We are in the process of incorporating the proposed method into the existing BMA-CRM software.

Appendix Proof of Lemma 2 in Chapter 2

Lemma 2 If $\boldsymbol{p}$ and $\boldsymbol{p}'$ are two skeletons generated by the method of Lee&Cheung. For these two skeletons, only initial guesses of location for the MTD are different, then $\boldsymbol{p} \sim \boldsymbol{p}'$.

Proof: Assume $v_1 < v_2$, without loss of generality, let $v_2 = v_1 + 1$.

According to the algorithm by Lee & Cheung (2009) in their paper, for v_1,

$$p_{i+1} = exp(\frac{log(p_T + \delta)log(p_i)}{log(P_T - \delta)}), i = v_1, \cdots, k-1$$

$$p_{v_1} = p_T$$

$$p_{i-1} = exp(\frac{log(p_T - \delta)log(p_i)}{log(P_T + \delta)}), i = 2, \cdots, v_1$$

For v_2:

$$p^{\star}_{i+1} = exp(\frac{log(p_T + \delta)log(p^{\star}_i)}{log(P_T - \delta)}), i = v_1 + 1, \cdots, k$$

$$p^{\star}_{v_1+1} = p_T$$

$$p^{\star}_{i-1} = exp(\frac{log(p_T - \delta)log(p_i)}{log(P_T + \delta)}), i = 2, \cdots, v_1 + 1$$

Given any k, if $k \in [v_1, \cdots, k-1]$, without loss of generality, let $k = v_1$, we have

$$\frac{log(p_{k+1})}{log(p^{\star}_{k+1})} = \frac{\frac{log(p_T+\delta)log(p_k)}{log(p_T-\delta)}}{\frac{log(p_T+\delta)log(p^{\star}_k)}{log(p_T-\delta)}} = \frac{log(p_{v_1})}{log(p^{\star}_{v_1})} = \frac{log(p_{v_1})}{log(p_{v_1-1})} = \frac{log(p_{v_1})}{\frac{log(p_T-\delta)log(p_{v_1})}{log(p_T+\delta)}} = \frac{log(p_T+\delta)}{log(p_T-\delta)}$$

If $k \in [2, v_1]$, without loss of generality, let $k = v_1$, we have

$$\frac{log(p_{k-1})}{log(p^{\star}_{k-1})} = \frac{\frac{log(p_T-\delta)log(p_{v_1})}{log(p_T+\delta)}}{\frac{log(p_T-\delta)log(p^{\star}_{v_1})}{log(p_T+\delta)}} \frac{log(p_{v_1})}{log(p^{\star}_{v_1})} = \frac{log(p_T+\delta)}{log(p_T-\delta)}$$

Thus, for any two skeletons generated by the method of LL & Cheung, the ratio of logarithm of the two skeleton is constant, that is, the two skeletons are equivalent based on the Definition 1 proposed in the paper, irrespective of prior guess of the MTD location.

Table 2.1: Simulation Results

Skeletons		Dose level 1	2	3	4	5	6	7	8	Avg # of toxicity
					Scenario 1					
	Tox rate	0.06	0.15	0.30	0.55	0.60	0.65	0.68	0.70	
Empirical	Sel %	0.8	15.9	**66.5**	15.7	0.9	0.1	0	0	8.7
	# Pts	4.0	6.8	**12.8**	5.3	0.9	0.1	0	0	
Optimal	Sel %	0.2	16.5	**72.1**	10.5	0.5	0	0	0.1	8.5
	# Pts	4.0	6.9	**13.2**	5.2	0.7	0	0	0	
Minimax	Sel %	0	15.4	**70.3**	13.5	0.6	0	0	0	8.5
	# Pts	3.9	6.4	**13.5**	5.4	0.7	0.1	0	0	
CRM	Sel %	0	13	**75.8**	10.4	0.3	0	0	0	7.8
	# Pts	3.8	7.7	**14.6**	3.5	0.2	0	0	0	
					Scenario 2					
	Tox rate	0.02	0.03	0.05	0.07	0.30	0.50	0.70	0.80	
Empirical	Sel %	0	0	0	10.1	**61.5**	26.3	1.9	0.2	7.2
	# Pts	3.2	3.0	3.1	4.4	**9.1**	6.1	1.0	0	
Optimal	Sel %	0	0	0	9.5	**71.7**	17.1	1.1	0.6	6.7
	# Pts	3.2	3.0	3.2	4.8	**10.1**	4.9	0.8	0	
Minimax	Sel %	0	0	0.2	12.1	**64.6**	21.0	1.4	0.7	6.8
	# Pts	3.2	3.0	3.1	4.7	**9.7**	5.2	0.9	0.1	
CRM	Sel %	0	0	0	16	**68**	14	0	0	5.4
	# Pts	3.2	3.3	4.4	5.9	**10.2**	2.8	0.2	0	
					Scenario 3					
	Tox rate	0.02	0.03	0.05	0.06	0.07	0.09	0.10	0.30	
Empirical	Sel %	0	0	0	0.5	1.0	2.7	17.4	**78.4**	3.4
	# Pts	3.2	3.1	3.2	3.3	3.4	3.6	4	**6.4**	
Optimal	Sel %	0	0	0	0	0.4	1.6	9.3	**88.7**	3.6
	# Pts	3.2	3.0	3.1	3.2	3.2	3.3	3.5	**7.5**	
Minimax	Sel %	0	0	0	0	1.0	3.7	14.3	**81.0**	3.3
	# Pts	3.2	3.0	3.1	3.3	3.4	3.6	3.9	**6.4**	
CRM	Sel %	0	0	0	10.2	14.8	19.2	22.5	**32.4**	2.4
	# Pts	3.2	3.3	4.4	4.9	4.9	3.9	3.0	**2.5**	
					Scenario 4					
	Tox rate	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	
Empirical	Sel %	2.7	21.8	**41.4**	27.3	5.8	0.6	0.1	0	8.3
	# Pts	5.6	7.5	**8.9**	5.6	1.9	0.4	0	0	
Optimal	Sel %	1.8	20.1	**49.7**	23.1	4.0	0.7	0	0.2	8.1
	# Pts	5.4	7.4	**9.7**	5.6	1.6	0.3	0	0	
Minimax	Sel %	1.8	20.1	**49.7**	23.1	4.0	0.7	0	0.2	8.1
	# Pts	5.4	7.4	**9.7**	5.6	1.6	0.3	0	0	
CRM	Sel %	1.3	26.6	**53**	17.7	1.3	0.1	0	0	7.5
	# Pts	4.9	9.9	**11.3**	3.4	0.4	0	0	0	

Table 1 (continued)

Skeletons		Dose level 1	2	3	4	5	6	7	8	Avg # of toxicity
					Scenario 5					
	Tox rate	0.20	0.30	0.40	0.50	0.60	0.65	0.70	0.75	
Empirical	Sel %	22.7	**40.6**	26.7	4.9	0.5	0	0	0	8.9
	# Pts	10.4	**9.4**	6.8	2	0.3	0	0	0	
Optimal	Sel %	21.3	**46.4**	20.7	5.1	0.4	0.5	0.2	0	8.7
	# Pts	11.0	**9.6**	5.9	1.9	0.4	0.1	0	0	
Minimax	Sel %	23.1	**43.0**	23.3	4.9	0.4	0.1	0	0	8.7
	# Pts	11.6	**9.1**	6.1	1.9	0.3	0	0	0	
CRM	Sel %	18.9	**53.9**	24.4	2.7	0.1	0	0	0	8.8
	# Pts	10.1	**12.9**	6.1	0.8	0	0	0	0	
					Scenario 6					
	Tox rate	0.02	0.06	0.08	0.12	0.20	0.30	0.40	0.50	
Empirical	Sel %	0	0	0.2	4.1	25.9	**38.8**	23.4	7.6	5.8
	# Pts	3.2	3.1	3.3	4.2	6.2	**6.1**	3.1	0.8	
Optimal	Sel %	0	0	0.2	5.4	27.5	**40.3**	17.1	9.5	5.5
	# Pts	3.3	3.1	3.5	4.6	6.4	**5.8**	2.4	0.9	
Minimax	Sel %	0	0	0.4	5.3	29.5	**37.0**	15.6	12.2	5.6
	# Pts	3.2	3.1	3.4	4.6	6.6	**5.5**	2.7	0.9	
CRM	Sel %	0	0	0.4	22.5	41.9	**23.4**	6.9	1.1	4.2
	# Pts	3.2	3.7	5.5	6.8	6.6	**3.3**	0.8	0.1	
					Scenario 7					
	Tox rate	0.02	0.03	0.04	0.06	0.08	0.10	0.30	0.50	
Empirical	Sel %	0	0	0	0	1.1	17	**50.9**	31	4.9
	# Pts	3.2	3.0	3.1	3.2	3.5	4.4	**6.1**	3.5	
Optimal	Sel %	0	0	0	0	2.3	16.5	**53.4**	27.8	4.6
	# Pts	3.2	3	3.1	3.3	3.6	4.9	**6.5**	2.4	
Minimax	Sel %	0	0	0	0.1	1.9	18.6	**48.3**	31.1	4.7
	# Pts	3.2	3.0	3.1	3.3	3.5	4.6	**5.8**	3.5	
CRM	Sel %	0	0	0.5	10	15.1	25.1	**41.7**	7.5	3.1
	# Pts	3.2	3.3	4.2	4.8	5.0	4.6	**3.9**	0.9	
					Scenario 8					
	Tox rate	0.03	0.07	0.10	0.15	0.20	0.30	0.50	0.70	
Empirical	Sel %	0	0	0.8	6.4	28	**45.3**	17.9	1.6	6.1
	# Pts	3.3	3.2	3.7	4.6	5.9	**5.9**	2.9	0.4	
Optimal	Sel %	0	0	1.1	9.2	31.1	**43.3**	10.1	5.2	5.8
	# Pts	3.3	3.2	3.7	4.9	6.4	**5.9**	2.2	0.4	
Minimax	Sel %	0	0.1	1.7	8.8	33.4	**37.0**	11.9	7.1	5.8
	# Pts	3.3	3.2	3.7	4.9	6.4	**5.9**	2.2	0.4	
CRM	Sel %	0	0	8.6	30.5	35.1	**21.0**	4.4	0	4.4
	# Pts	3.3	4.1	6.4	7.3	5.5	**2.7**	0.6	0.1	

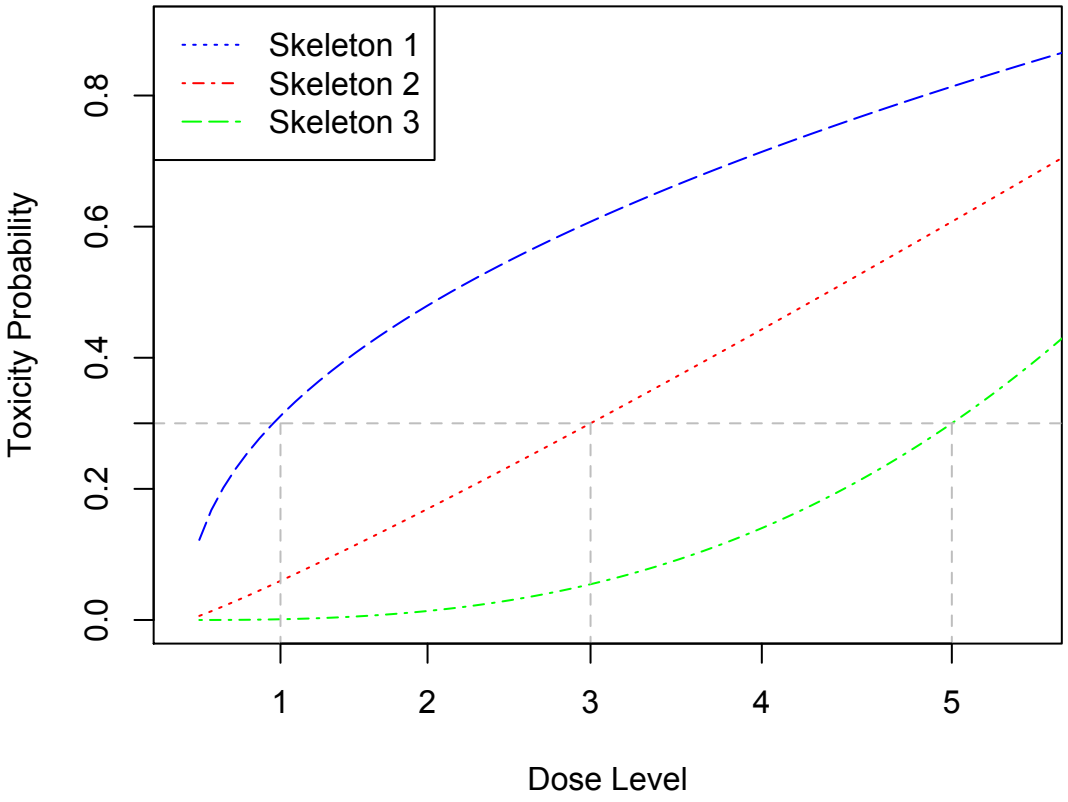


Figure 2.1.: Example of equivalent skeletons for the CRM

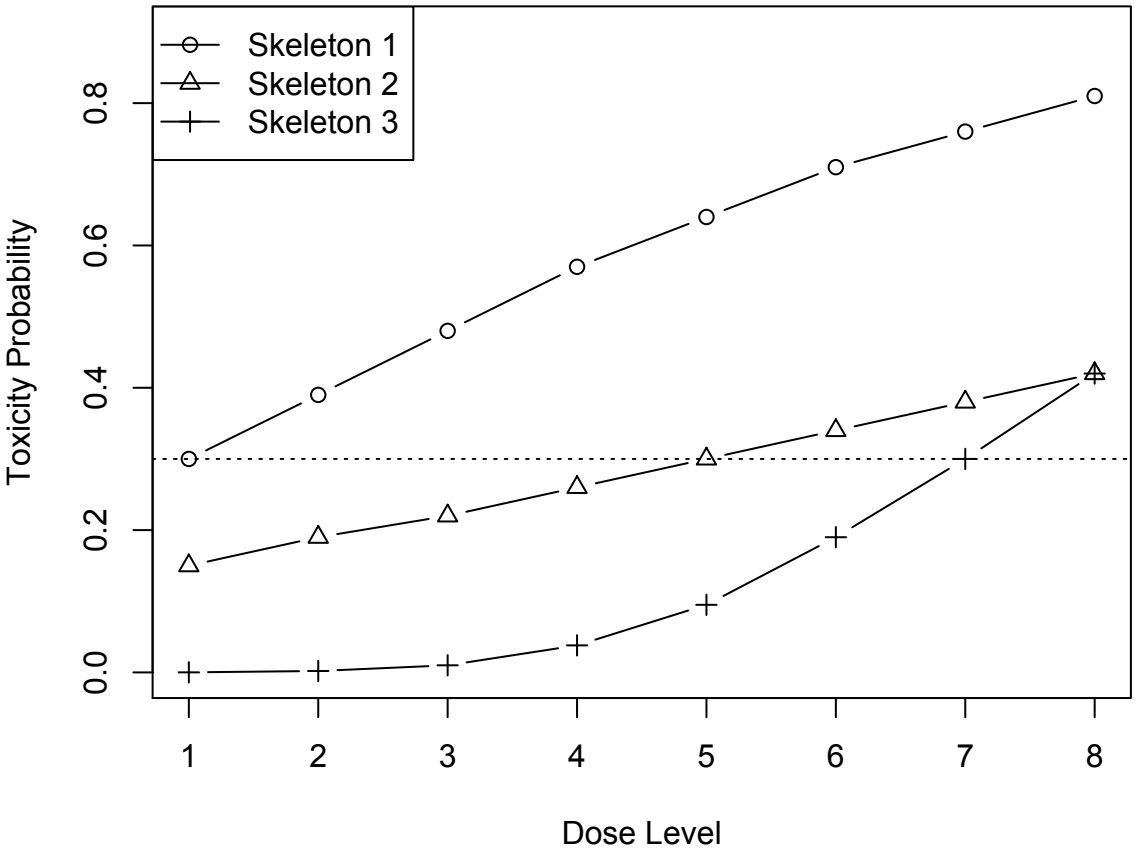


Figure 2.2.: Three empirically chosen skeletons

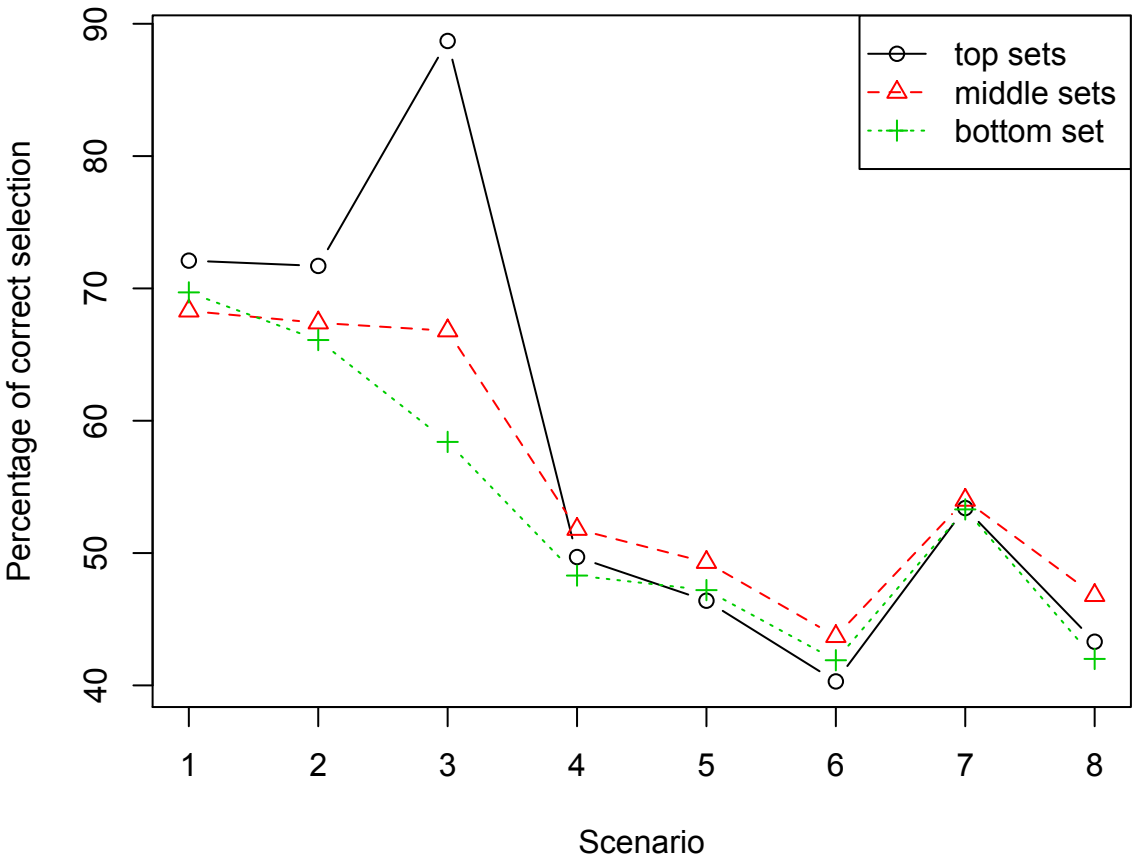


Figure 2.3.: The percentage of correct dose selection (PCS) of the MTD in eight scenarios based on the top, middle and bottom 20 skeleton sets.

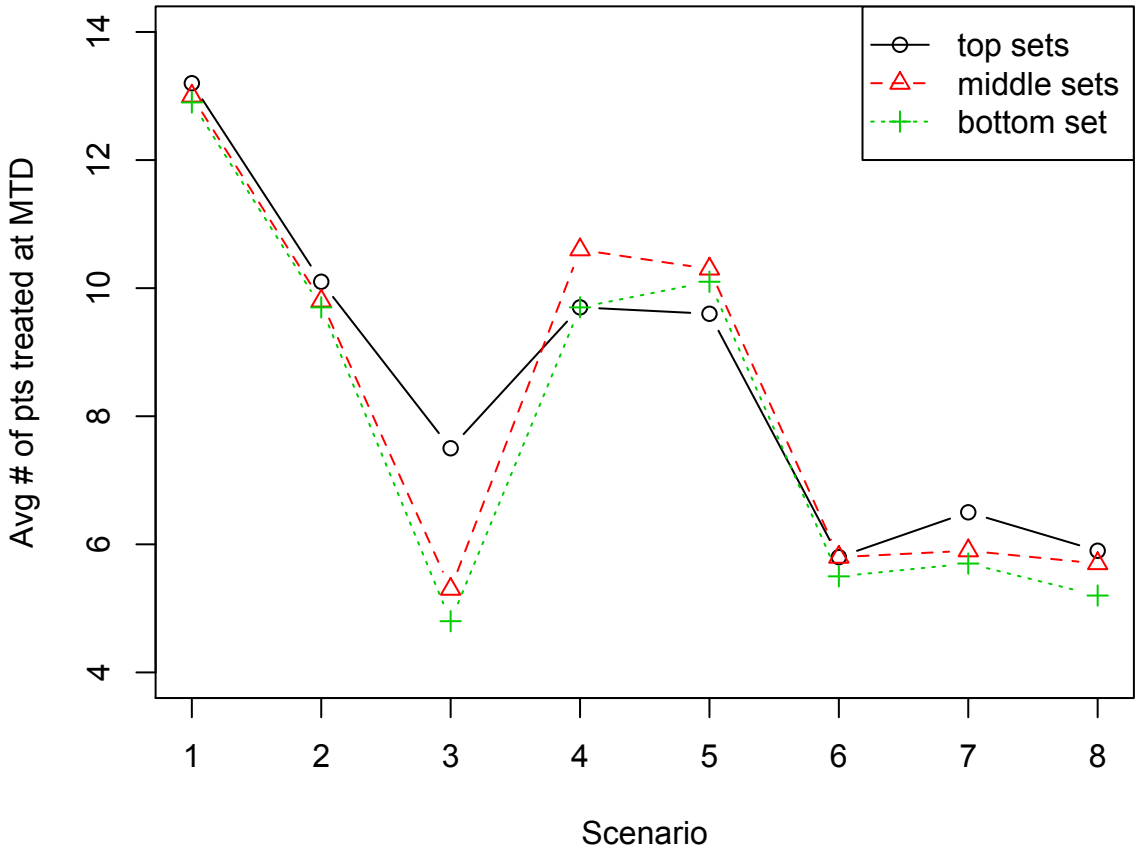


Figure 2.4.: The average number of patients treated at the MTD in eight scenarios based on the top, middle and bottom 20 skeleton sets.

Table 2.2: Simulation results when the recommended skeleton set selected from skeleton sets with top, middle and bottom 20 Q values

Q value		Dose level 1	2	3	4	5	6	7	8	Avg # of toxicity
					Scenario 1					
	Tox rate	0.06	0.15	0.30	0.55	0.60	0.65	0.68	0.70	
Top	Sel %	0.2	16.5	**72.1**	10.5	0.5	0	0	0.1	8.5
	# Pts	4.0	6.9	**13.2**	5.2	0.7	0	0	0	
Middle	Sel %	0.1	15.3	**68.3**	15.1	0.9	0.2	0	0	8.6
	# Pts	4.0	6.5	**13.0**	5.6	0.8	0.1	0	0	
Bottom	Sel %	0.3	17.2	**69.7**	11.7	0.9	0	0	0	8.6
	# Pts	4.1	6.9	**12.9**	5.3	0.6	0.1	0	0	
					Scenario 2					
		0.02	0.03	0.05	0.07	0.30	0.50	0.70	0.80	
Top	Sel %	0	0	0	9.5	**71.7**	17.1	1.1	0.6	6.7
	# Pts	3.2	3.0	3.2	4.8	**10.1**	4.9	0.8	0	
Middle	Sel %	0	0	0	8.2	**67.4**	23.5	0.9	0	6.7
	# Pts	3.3	3.1	3.2	4.7	**9.8**	5.1	0.8	0	
Bottom	Sel %	0	0	0.3	9.0	**66.1**	23.3	1.3	0	6.7
	# Pts	3.2	3.0	3.2	4.8	**9.7**	5.2	0.8	0	
					Scenario 3					
		0.02	0.03	0.05	0.06	0.07	0.09	0.10	0.30	
Top	Sel %	0	0	0	0	0.4	1.6	9.3	**88.7**	3.6
	# Pts	3.2	3.0	3.1	3.2	3.2	3.3	3.5	**7.5**	
Middle	Sel %	0	0	0	0.6	1.2	5.9	25.5	**66.8**	3.2
	# Pts	3.2	3.0	3.2	3.4	3.6	3.9	4.4	**5.3**	
Bottom	Sel %	0	0	0.1	1.2	3.3	11.8	25.2	**58.4**	3.2
	# Pts	3.2	3.0	3.2	3.7	3.9	4.1	3.9	**4.8**	
					Scenario 4					
		0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	
Top	Sel %	1.8	20.1	**49.7**	23.1	4.0	0.7	0	0.2	8.1
	# Pts	5.4	7.4	**9.7**	5.6	1.6	0.3	0	0	
Middle	Sel %	1.1	19.8	**51.8**	22.4	4.3	0.3	0	0	8.0
	# Pts	5.1	7.5	**10.6**	5.3	1.3	0.2	0	0	
Bottom	Sel %	1.7	21.8	**48.3**	22.9	4.9	0.4	0	0	8.1
	# Pts	5.1	7.9	**9.7**	5.5	1.5	0.2	0	0	

Table 2 (continued)

Q value		Dose level 1	2	3	4	5	6	7	8	Avg # of toxicity
					Scenario 5					
		0.20	0.30	0.40	0.50	0.60	0.65	0.70	0.75	
Top	Sel %	21.3	**46.4**	20.7	5.1	0.4	0.5	0.2	0	8.7
	# Pts	11.0	**9.6**	5.9	1.9	0.4	0.1	0	0	
Middle	Sel %	20.5	**49.3**	23.4	3.6	0.1	0	0	0	8.7
	# Pts	11.3	**10.3**	5.9	1.5	0.3	0	0	0	
Bottom	Sel %	21.7	**47.2**	22.7	3.7	0.3	0	0	0	8.7
	# Pts	10.6	**10.1**	6.4	1.7	0.3	0	0	0	
					Scenario 6					
		0.02	0.06	0.08	0.12	0.20	0.30	0.40	0.50	
Top	Sel %	0	0	0.2	5.4	27.5	**40.3**	17.1	9.5	5.5
	# Pts	3.3	3.1	3.5	4.6	6.4	**5.8**	2.4	0.9	
Middle	Sel %	0	0.1	0.2	5.3	27.0	**43.7**	19.2	4.5	5.4
	# Pts	3.3	3.1	3.4	4.6	6.5	**5.8**	2.7	0.7	
Bottom	Sel %	0	0	0.3	6.2	32.9	**41.9**	15.3	3.4	5.3
	# Pts	3.2	3.1	3.5	4.9	7.0	**5.5**	2.2	0.5	
					Scenario 7					
		0.02	0.03	0.04	0.06	0.08	0.10	0.30	0.50	
Top	Sel %	0	0	0	0	2.3	16.5	**53.4**	27.8	4.6
	# Pts	3.2	3	3.1	3.3	3.6	4.9	**6.5**	2.4	
Middle	Sel %	0	0	0	0.3	3.0	18.0	**54.0**	24.7	4.4
	# Pts	3.2	3.1	3.1	3.4	3.7	4.9	**5.9**	2.8	
Bottom	Sel %	0	0	0	0.6	2.9	22.2	**53.3**	21.0	4.2
	# Pts	3.2	3.0	3.1	3.5	4.0	4.9	**5.7**	2.6	
					Scenario 8					
		0.03	0.07	0.10	0.15	0.20	0.30	0.50	0.70	
Top	Sel %	0	0	1.1	9.2	31.1	**43.3**	10.1	5.2	5.8
	# Pts	3.3	3.2	3.7	4.9	6.4	**5.9**	2.2	0.4	
Middle	Sel %	0	0	1.4	10.4	28.8	**46.8**	12.1	0.5	5.6
	# Pts	3.4	3.2	3.7	5.4	6.5	**5.7**	1.9	0.2	
Bottom	Sel %	0	0	1.8	11.9	32.6	**42.0**	11.2	0.5	5.5
	# Pts	3.4	3.2	4.0	5.6	6.7	**5.2**	1.7	0.2	

3. Bridging continual reassessment method for phase I clinical trials in different ethnic populations

3.1 Introduction

This chapter is based on "Bridging continual reassessment method for phase I clinical trials in different ethnic populations" published in Statistics in Medicine (2015) [81] coauthored with Suyu Liu, Jielai Xia, Qing Huang and Ying Yuan. Permission from the journal has been granted for use in conjunction with the thesis.

As introduced in Chapter 1, traditionally, phase I trials are conducted in a "one-size-fits-all" fashion. That is, once the MTD of a drug is established in a landmark study based on a certain ethnic population (e.g., a Caucasian population), the results are directly extrapolated to other ethnic populations (e.g., an Asian population). Unfortunately, accumulating evidence shows that such a one-size-fits-all dose-finding paradigm is problematic, and ethnicity plays an important role in a patient's response to a drug [21]. The inter-ethnic differences have being recognized by drug regulatory authorities. In 1999, the FDA published the guidance, "Ethnic Factors in the Acceptability of Foreign Clinical Data" (i.e., E5 Guidance), which suggested the need to distinguish three ethnic categories (i.e., Asian, Black, and Caucasian) in drug development. The guidance identified situations for which drugs could be ethnically sensitive and suggested the types of bridge studies that may be required to extrapolate clinical trial results from one region to another.

The goal of this chapter is to address the following bridge trial design question: given that a landmark phase I trial has been conducted in a landmark population and the corresponding MTD has been established, how do we design a follow-up trial

(i.e., a bridge trial) to find the MTD for a new population. To address this issue, we propose the bridging CRM (B-CRM) design, which utilizes the dose-toxicity data obtained from the landmark trial to achieve efficient dose finding in the follow-up trial, while also acknowledging inter-ethnic heterogeneity.

The remainder of the chapter is organized as follows. In Section 3.2, after a very brief review of the original CRM again, we propose a novel mixture estimate of the dose-toxicity curve using the landmark trial data. Based on this estimate, we present the procedure of using multiple skeletons to accommodate the inter-ethnic heterogeneity and that of using Bayesian model averaging to make the decision of dose assignment. In Section 3.3, we investigate the operating characteristics of the proposed B-CRM using simulation studies. In Section 3.4, we illustrate the proposed design using a phase I clinical trial for advanced solid tumors, and conclude with a brief discussion in Section 3.5.

3.2 Methods

3.2.1 Continual Reassessment Method

Let $(d_1, \ldots, d_J)$ denote a set of J prespecified doses for the drug under investigation. We assume that the dose-limiting toxicity (DLT) is recorded as a binary outcome and the true dose toxicity monotonically increases with respect to the dose level. Let $(p_1, \ldots, p_J)$ be the prespecified toxicity probabilities of the J doses, $p_1 < \ldots < p_J$, which are also known as the skeleton. The CRM links the true toxicity probability at dose d_j, denoted as $\pi(d_j)$, with the prespecified prior toxicity probability p_j, using a working dose-toxicity model, such as

$$\pi(d_j) = p_j^{\exp(\alpha)} \tag{3.1}$$

for $j = 1, \ldots, J$, where α is an unknown parameter. To conduct the trial, the CRM continuously updates the estimate of the dose-toxicity model using the accrued information, and adaptively assigns incoming patients to the dose with an estimated toxicity probability closest to the prespecified target toxicity probability, ϕ. Once the maximum sample size is reached, the dose with a posterior toxicity probability closest to ϕ is selected as the MTD.

One important feature of the CRM is that the prior information on the dose-toxicity curve can be naturally incorporated into the model through the specification of the skeleton, which enhances the performance of the design. In the proposed B-CRM, we exploit this feature of the CRM to borrow the dose-toxicity information from the landmark trial for finding the MTD in the follow-up trial. In the next subsection, we first describe a method to estimate the dose-toxicity curve based on the dose-toxicity data generated by the landmark trial, and then discuss how to incorporate such rich prior information into the follow-up trial by specifying multiple skeletons.

3.2.2 Estimation of dose-toxicity curve in landmark population

We assume that a landmark phase I trial has been conducted in a (landmark) population, with J_L prespecified doses, $b_1 < \ldots < b_{J_L}$. The trial identified dose b_{j^*} as the MTD and resulted in binomial data $\mathcal{D}_L = (x_j, m_j)$, where x_j is the number of patients who experienced toxicity, and m_j is the total number of patients treated at dose b_j, for $j = 1, \ldots, J_L$. Given $\mathcal{D}_L$, a straightforward way to estimate the dose-toxicity curve is to fit a probit model

$$\pi_j^{(P)} \equiv \pi^{(P)}(b_j) = \Phi(\beta_0 + \beta_1 b_j), \tag{3.2}$$

where the superscript in $\pi_j^{(P)}$ indicates it is a parametric estimate; $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution; β_0 and β_1 are intercept

and slope parameters, respectively. We require $\beta_1 > 0$ such that toxicity monotonically increases with the dose. We adopt the probit model because of its intuitive toxicity tolerance interpretation [36]. The tolerance is defined as the dose intensity level below which toxicity does not occur and above which toxicity occurs. If we assume that the tolerance varies from subject to subject and is normally distributed, then the dose-toxicity curve follows a probit model of the form given by (3.2). This parametric approach is simple, but when the model is misspecified, the resulting estimate may be severely biased. Alternatively, we can nonparametrically estimate the toxicity probability of dose b_j using isotonic regression [37, 38],

$$\hat{\pi}_j^{(NP)} = \max_{0 \le u \le j} \min_{j \le v \le J_L} \frac{\sum_{k=u}^{v} x_k}{\sum_{k=u}^{v} m_k}.$$

This isotonic estimate satisfies the monotonic constraint $\pi_1^{(NP)} \le \ldots \le \pi_{J_L}^{(NP)}$, and can be easily obtained by applying the pooled–adjacent-violators algorithm (PAVA) [39] to the observed toxicity rate $\gamma_j = x_j/m_j$, $j = 1, \ldots, J_L$. Operatively, the PAVA replaces any adjacent γ_j's that violate the nondecreasing order by their (weighted) average so that the resulting estimates, $\hat{\pi}_j^{(NP)}$, become monotonic. Bhattacharya and Kong [38] showed that $\hat{\pi}_j^{(NP)}$ is consistent under mild conditions. The drawback of isotonic regression is that the resulting estimates can be highly variable, and it is difficult to estimate the toxicity probabilities for doses outside the observed dose range, $[b_1, b_{J_L}]$.

To inherit the merits of parametric regression and nonparametric isotonic regression, we propose a mixture (or weighted average) estimator of toxicity probabilities,

$$\hat{\pi}_j = w_j \hat{\pi}_j^{(P)} + (1 - w_j)\hat{\pi}_j^{(NP)}, \tag{3.3}$$

where weight w_i is chosen in a data-driven way such that if $\hat{\pi}_j^{(P)}$ is more accurate, more weight is assigned to the probit regression, and if $\hat{\pi}_j^{(NP)}$ is more accurate, more

weight is assigned to the isotonic regression. We propose the following data-based weight,

$$w_j = \frac{\lambda_j}{\lambda_j + 1},$$

where

$$\lambda_j = \frac{\left(\hat{\pi}_j^{(P)}\right)^{x_j} \left(1 - \hat{\pi}_j^{(P)}\right)^{m_j - x_j}}{\left(\hat{\pi}_j^{(NP)}\right)^{x_j} \left(1 - \hat{\pi}_j^{(NP)}\right)^{m_j - x_j}} \tag{3.4}$$

is the (estimated) likelihood ratio evaluated at dose level j under the probit model and isotonic regression. Using the likelihood ratio as a weight, the parametric or nonparametric estimate that fits the data better will receive a higher weight. As a result, the proposed mixture estimator has the consistency property described in the following theorem (see the Appendix for the proof).

Theorem 1. The proposed mixture estimator in (3.3) is a consistent estimate of π_j.

Although $\{\hat{\pi}_j^{(P)}\}$ and $\{\hat{\pi}_j^{(NP)}\}$ are both monotonic, as weight w_j varies across doses, the mixture estimator $\{\hat{\pi}_j\}$ may occasionally violate the monotonicity assumption in finite samples. If that occurs, we can apply the PAVA algorithm to the $\hat{\pi}_j$'s to impose monotonicity. The transformed estimate will not take a form exactly same as that given by (3.3), but it remains to be a consistent estimate of π_j. For an arbitrary dosage d between b_j and b_{j+1}, its toxicity probability can be estimated using linear interpolation

$$\hat{\pi}(d) = \hat{\pi}_j + \frac{d - b_j}{b_{j+1} - b_j}(\hat{\pi}_{j+1} - \hat{\pi}_j), \qquad b_j \le d \le b_{j+1}.$$

Other more sophisticated methods, such as smoothing splines, can also be used to extrapolate the $\hat{\pi}_j$'s and obtain the estimate of $\pi(d)$. However, the simple linear interpolation is typically adequate because our goal here is not to pursue a precise

estimate of the entire dose-response curve for the landmark population, but to utilize the landmark trial data to provide a ballpark estimate of the toxicity probabilities at some prespecified doses (i.e., skeleton) for the follow-up population to facilitate the dose finding for the follow-up trial. Actually, as we never observe any data between b_j and b_{j+1}, all extrapolation methods are based on certain untestable model assumptions, and the observed data cannot inform us as to which extrapolation method is better. In addition, similar to the standard CRM, under the proposed B-CRM described below, the accumulating data collected in the follow-up trial will quickly dominate the skeleton and limit the influence of the skeleton.

3.2.3 Bridging CRM

We now consider how to design the follow-up trial to find the MTD for a new population, given that the estimate of the dose-toxicity curve has been obtained from the landmark trial as described above. We assume that a set of J doses, $d_1 < \ldots < d_J$, are under investigation in the follow-up trial. These doses do not have to be the same set of doses previously studied in the landmark trial (i.e., $b_1, \ldots, b_{J_L}$), and can be chosen based on the data collected from the landmark trial. For example, based on the dose-toxicity estimate (3) obtained from the landmark trial, we can choose the d_j's locally to the MTD. This can be done by taking the MTD identified in the landmark trial as the middle dose, and then determining other candidate doses by backsolving the dose-toxicity function such that the estimated toxicity probabilities of the other doses are equally spaced around the MTD. In general, no matter which approach is taken, the MTD identified in the landmark trial, i.e., b_{j^*}, should be included as one of the investigational doses in the follow-up trial. Without a loss of generality, we assume that this dose corresponds to dose level t in the follow-up trial (i.e., $d_t = b_{j^*}$).

Because of inter-ethnic heterogeneity, we do not expect the dose-toxicity curve for the new population to be exactly the same as that of the landmark population.

On the other hand, we also do not expect that these two curves dramatically differ from each other because they concern the same drug used to treat patients with the same type of disease. In practical use, the dose-toxicity curve of the new population should mostly resemble that of the landmark population, with some deviation. As a result, the MTD for the new population should be in the neighborhood of that of the landmark population (e.g., one dose level difference). Let $\hat{p}_j = \hat{\pi}(d_j)$, $j = 1, \ldots, J$, denote the estimate of toxicity probability of d_j based on the landmark trial data. Under the CRM framework, we incorporate such prior information using three sets of skeletons:

$$
\begin{aligned}
&\text{skeleton 1}: \quad && p_j = \hat{p}_j \\
&\text{skeleton 2}: \quad && p_j = \hat{p}_{j+1} \quad \text{for } j = 1, \ldots, J-1, \\
& && p_J = \frac{\hat{p}_J + 1}{2} \\
&\text{skeleton 3}: \quad && p_j = \hat{p}_{j-1} \quad \text{for } j = 2, \ldots, J, \\
& && p_1 = \frac{\hat{p}_1}{2}.
\end{aligned}
$$

That is, skeleton 1 represents that *a priori* the new population has the same toxicity profile as that of the landmark population; whereas skeletons 2 and 3 shift the dose-toxicity curve one level up and one level down, respectively. Under skeletons 1, 2 and 3, the MTD for the new population is *a priori* the same as, one level lower or one level higher than that for the landmark population. For skeleton 2, when we shift the toxicity probabilities one level up, the toxicity probability of the highest dose level (i.e., p_J) will move out of the existing range, and thus we take p_J as the middle value between $\hat{p}_J$ and 1. Similarly, we set the toxicity probability of the lowest dose level (i.e., p_1) as the middle value between 0 and $\hat{p}_1$ in skeleton 3 when shifting the toxicity probability one level down. Following Yin and Yuan [12], we regard each skeleton as a CRM model and use the Bayesian model averaging (BMA) [40, 41] approach to

estimate toxicity probabilities across multiple skeletons for adaptive dose assignment and selection.

Specifically, let $(M_1, \ldots, M_K)$ be the models corresponding to each set of prior guesses of the toxicity probabilities $\{(p_{11}, \ldots, p_{1J}), \ldots, (p_{K1}, \ldots, p_{KJ})\}$, where $K = 3$. Model M_k $(k = 1, \ldots, K)$ in the CRM is given by

$$\pi_{kj}(\alpha_k) = p_{kj}^{\exp(\alpha_k)}, \qquad j = 1, \ldots, J,$$

which is based on the kth skeleton $(p_{k1}, \ldots, p_{kJ})$. Let $\text{pr}(M_k)$ be the prior probability that model M_k is the true model, i.e., the probability that the kth skeleton $(p_{k1}, \ldots, p_{kJ})$ matches the true dose-toxicity curve. The value of $\text{pr}(M_k)$ should reflect the prior knowledge of whether the new population is likely to be less or more tolerable to the drug. For example, Asians are often expected to be less tolerable to certain drugs than Caucasians. Thus, if the landmark population is Caucasian, we may assign a high prior probability to skeleton 2 and a low prior probability to skeleton 3 when the new population is Asian. When there is no prior information regarding the relative tolerance between landmark and new populations, we can assign equal weights to the different skeletons by simply setting $\text{pr}(M_k) = 1/K$. Suppose at a certain stage of the trial, among n_j patients treated at dose level j, y_j patients have experienced DLT. Let $\mathcal{D} = \{(n_j, y_j), j = 1, \ldots, J\}$ denote the observed data, the likelihood function under model M_k is

$$L(\mathcal{D}|\alpha_k, M_k) = \prod_{j=1}^{J} \{p_{kj}^{\exp(\alpha_k)}\}^{y_j} \{1 - p_{kj}^{\exp(\alpha_k)}\}^{n_j - y_j}.$$

The posterior model probability for M_k is given by

$$\text{pr}(M_k|\mathcal{D}) = \frac{L(\mathcal{D}|M_k)\text{pr}(M_k)}{\sum_{i=1}^{K} L(\mathcal{D}|M_i)\text{pr}(M_i)},$$

where $L(\mathcal{D}|M_k)$ is the marginal likelihood of model M_k,

$$L(\mathcal{D}|M_k) = \int L(\mathcal{D}|\alpha_k, M_k) f(\alpha_k|M_k) d\alpha_k,$$

α_k is the power parameter in the CRM associated with model M_k, and $f(\alpha_k|M_k)$ is the prior distribution of α_k under model M_k, e.g., $f(\alpha_k|M_k) \sim N(0, 2)$. The BMA estimate for the toxicity probability at each dose level is given by

$$\bar{\pi}_j = \sum_{k=1}^{K} \hat{\pi}_{kj} \text{pr}(M_k|\mathcal{D}), \quad j = 1, \ldots, J, \tag{3.5}$$

where $\hat{\pi}_{kj}$ is the posterior mean of the toxicity probability of dose level j under model M_k, i.e.,

$$\hat{\pi}_{kj} = \int p_{kj}^{\exp(\alpha_k)} \frac{L(\mathcal{D}|\alpha_k, M_k) f(\alpha_k|M_k)}{\int L(\mathcal{D}|\alpha_k, M_k) f(\alpha_k|M_k) d\alpha_k} d\alpha_k.$$

Alternatively, we can use the model selection approach to estimate the toxicity probabilities and make the decision of dose assignment. That is, at each point of decision making for dose assignment, we select the model with the highest posterior probability, i.e., model $k^* = \text{argmax}_{k \in (1,\ldots,K)}(\text{pr}(M_k|\mathcal{D}))$, and use that model to make inference and dose assignment. However, our numerical study shows that the BMA approach performs slightly better than the model selection approach (results not shown), thus we focus on the BMA approach hereafter.

3.2.4 Dose-finding Algorithm

The dose-finding algorithm for our B-CRM design is described as follows,

1. Patients in the first cohort are treated at dose d_{t-1}, i.e., the dose one level lower than the MTD identified in the landmark trial (i.e., d_t). Note that we choose d_{t-1}, rather than d_t, as the starting dose to fit the physician's inclination to be conservative for safety purposes.

2. At the current dose level, j^{curr}, based on the observed data, we calculate the BMA estimates for the toxicity probabilities, $\bar{\pi}_j$ $(j = 1, \ldots, J)$, and identify dose level j^{best} that has a toxicity probability closest to ϕ, i.e.,

$$j^{\text{best}} = \text{argmin}_{j \in (1,\ldots,J)} |\bar{\pi}_j - \phi|.$$

If $j^{\text{curr}} > j^{\text{best}}$, we de-escalate the dose level to $j^{\text{curr}} - 1$; if $j^{\text{curr}} < j^{\text{best}}$, we escalate the dose level to $j^{\text{curr}} + 1$; otherwise, the dose stays at the same level as j^{curr} for the next cohort of patients. Being conservative, we restrict the dose change to one level at a time.

3. Once the maximum sample size is reached, the dose that has the toxicity probability closest to ϕ is selected as the MTD.

In addition, we impose the following safety stopping rule: if $\text{pr}(\pi_1 > \phi|\mathcal{D}) > 0.9$, the trial is terminated for safety. That is, if the lowest dose has a high probability of being overly toxic, we should stop the trial early for safety. The software to implement the proposed B-CRM design (written in R) can be found in Supplementary Materials, and also is available for free download at http://odin.mdacc.tmc.edu/∼yyuan/.

3.3 Simulation Studies

We investigated the operating characteristics of the proposed B-CRM design through simulation studies. We considered 6 dose levels with a target toxicity probability of $\phi = 30\%$. The maximum sample size for the follow-up trial was 21 patients in cohorts of size 3. Suppose that the landmark trial has been done using a certain phase I trial design (e.g., the "3+3" design), which yielded the following data: the number of patients treated at each dose $(m_1, \ldots, m_6) = (3, 3, 3, 6, 3, 0)$, and the number of patients who experienced dose-limiting toxicity at each dose $(x_1, \ldots, x_6) = (0, 0, 0, 1, 2, 0)$. Dose level 4 was identified as the MTD in the land-

mark trial. For the follow-up trial, we considered 6 toxicity scenarios that differ in the location of the true MTD. We note that the actual dosage of these 6 dose levels can be different from that in the landmark trial. In addition, in practice, the number of doses studied in the follow-up trial is not necessarily the same as that of the landmark trial. We compared the proposed B-CRM to two methods: the conventional CRM (for a stand-alone trial), which does not actively borrow information from the landmark trial, and the method proposed by Morita [35], which borrows information from the landmark trial using an informative prior under the CRM (referred to as the IP-CRM). For the IP-CRM method, a one-parameter logistic regression model was used,

$$\pi_j = \frac{\exp(\beta_0 + \beta_1 b_j)}{1 + \exp(\beta_0 + \beta_1 b_j)},$$

where $\beta_0 \equiv 3$, and dosage b_j was specified using "backward fitting" [42] such that the prior estimates of the toxicity probabilities match the estimates from the landmark trial, which are obtained by fitting a logistic model to the landmark trial data. Following Morita [35], we assumed that β_1 follows a gamma prior $Ga(5, 5)$. For the conventional CRM, we chose the skeleton (0.12, 0.20, 0.30, 0.40, 0.50, 0.6) based on the method of Lee and Cheung [43], assuming an indifference interval of 0.1. We used the same starting dose (i.e., dose level 3) for all three designs. Strictly speaking, because this starting dose is chosen based on the landmark trial data, the CRM we considered here actually borrowed some information from the landmark trial.

Table 3.1 shows the simulation results, including selection percentages and the number of patients treated at each dose, based on 1,000 simulated trials. In scenario 1, the MTD is the fourth dose, the same dose level as the MTD of the landmark trial. Compared to the CRM, the B-CRM yielded a 7.8% higher percentage of correct selection (PCS), and assigned about 2 more patients to the MTD. In addition, the B-CRM was also 9.6% less likely to select the overly toxic doses (i.e., dose levels 5 and 6) than the CRM. The IP-CRM performed best with the highest PCS. However, as we

will see, when the MTD of the follow-up trial differs from that of the landmark trial, the performance of the IP-CRM can be poor. Scenarios 2 and 3 respectively present the cases in which the MTD in the follow-up trial is one level higher or lower than that identified in the landmark population. In these cases, the B-CRM outperformed the CRM with 6%-8% higher PCS and also lower probabilities of selecting the overly toxic doses. The IP-CRM showed large variation: it performed reasonably well in scenario 3 (PCS=54.0%), but poorly in scenario 2 (PCS=36.8%). In scenarios 4 and 5, the MTD in the follow-up trial is two levels different from that in the landmark trial. The B-CRM consistently performed well, with the PCS ranging from 58.6% to 63.1%. The CRM and IP-CRM were less stable. The CRM performed very well in scenario 4, but not as well in scenario 5; whereas the IP-CRM performed well in scenario 5, but very poorly in scenario 4. Scenario 6 has the first dose as the MTD, which is three levels different from the MTD identified in the landmark trial. In this case, the B-CRM and IP-CRM yielded similar PCS, but the B-CRM was 18.1% less likely to select doses above the MTD. The PCS of the CRM is about 8% lower than those of the B-CRM and IP-CRM. As a sensitivity analysis, we also investigated the operating characteristics of the designs given a different set of landmark trial data, i.e., $(m_1, \ldots, m_6) = (3, 3, 6, 3, 0, 0)$ and $(x_1, \ldots, x_6) = (0, 0, 1, 2, 0, 0)$, for which dose level 3 was identified as the MTD of the landmark trial. The pattern of the results is generally similar to that given above (see Table 3.2).

As we noted previously, in practice, we do not expect the dose-toxicity curve for the new population to dramatically differ from that for the landmark population because they concern the same drug that is used to treat patients with the same type of disease. The MTD for the new population should be in the neighborhood of that of the landmark population, i.e., scenarios 1, 2 and 3 are more likely encountered in practice than other scenarios. In the case when there is strong prior knowledge that the MTD for the new population is much lower than the MTD for the landmark

population, when specifying the investigational doses for the follow-up trial, we should choose more dose levels that are lower than the MTD identified in the landmark trial. Specifically, we can choose the dose that is most likely to be the MTD (for the new population) as dose level 4 (of the follow-up trial), and then add other doses. By doing so, we ensure that the MTD for the follow-up trial is still in the neighborhood of that of the landmark trial in terms of dose level (i.e., dose level 4), although they may be very different in terms of actual dosage. One advantage of using the power model (3.1) is that the actual dosages are not directly used in the model; rather we use their associated toxicity probabilities (i.e., the skeleton). For example, in our simulation, we did not need to specify the actual dosages for the follow-up trials.

3.4 Application

A multi-center phase I study was recently conducted to find the MTD of BKM120 in adult patients with advanced solid tumors [44]. BKM120 is a potent, highly specific oral inhibitor of the intracellular phosphatidylinositol-3-kinase (PI3K) pathway, which regulates cellular functions, such as cell proliferation, growth, survival and apoptosis. Selective inhibition of the PI3K pathway provides a promising therapeutic approach to treat cancer. A total of six doses were investigated, i.e., 12.5, 25, 50, 80, 100, or 150 mg. Dose-limiting toxicities (DLTs) were evaluated during the first treatment cycle (28 days). The main DLTs were defined as any grade-3 or higher hematologic or nonhematologic toxicity according to the Common Terminology Criteria for Adverse Events (CTCAE), version 3.0. The MTD was defined as the highest dose of BKM120 yielding a DLT rate not higher than 33%. A total of 35 patients were treated in the trial. The resulting dose-toxicity data are shown in Table 3.3, with a dose of 100mg selected as the MTD. As the patients in this trial came entirely from the United States, Canada, the Netherlands and Spain, the identified MTD may not be applicable to Asian populations. Our collaborators at the Fourth Military Medical

University in China are interested in conducting a follow-up phase I bridge trial to establish the MTD of BKM120 in Chinese patients.

We applied the proposed B-CRM to design the follow-up trial, in which we considered the same six doses that were evaluated in the landmark trial. The maximum sample size was 24 patients. Based on the dose-toxicity data from the landmark trial (see Table 3.3), we estimated the toxicity probabilities of the doses using the proposed mixture estimator, yielding $(\hat{\pi}_1, \ldots, \hat{\pi}_6)$ =(0.002, 0.004, 0.014, 0.137, 0.220, 0.546). Accordingly, we constructed three skeletons:

$$\begin{aligned}
\text{skeleton } 1 : (p_1, \ldots, p_6) &= (0.002, 0.004, 0.014, 0.137, 0.220, 0.546) \\
\text{skeleton } 2 : (p_1, \ldots, p_6) &= (0.004, 0.014, 0.137, 0.220, 0.546, 0.773) \\
\text{skeleton } 3 : (p_1, \ldots, p_6) &= (0.001, 0.002, 0.004, 0.014, 0.137, 0.220),
\end{aligned}$$

where skeletons 1 to 3 represent that the MTD for Chinese patients is *a priori* the same as, one level lower or one level higher than that for non-Asian patients. To evaluate the operating characteristics of the B-CRM for this trial, we considered four scenarios that differ in both the location of the MTD and the shape of the true dose-toxicity curve for the follow-up trial (see Figure 3.1). We simulated 1,000 trials under each scenario. For the purpose of comparison, we also applied the CRM and IP-CRM. The results (see Table **??**) show that, compared to the CRM and IP-CRM, the proposed B-CRM has the most reliable operating characteristics. It selected the true MTD consistently with high probabilities (62.5% to 68.5%) and assigned the majority of the patients (i.e., about 10 or more) to the MTD. The CRM and IP-CRM performed well in some scenarios (e.g., scenario 2 for the CRM, and scenario 4 for the IP-CRM), but worse in other scenarios (e.g., scenarios 1 and 4 for the CRM, and scenario 2 for the IP-CRM).

3.5 Summary

We have proposed the B-CRM to find the MTD of a drug for a new ethnic population, given that a landmark trial has been conducted to establish the MTD in a landmark population. Our method borrows dose-toxicity information from the landmark trial and also accounts for inter-ethnic differences. We propose a novel mixture estimator to estimate the dose-toxicity curve using the data yielded by the landmark trial. Based on the resulting estimate, we form multiple skeletons to borrow information from the landmark trial and also accommodate the inter-ethnic heterogeneity. We use the Bayesian model averaging approach to make inference across multiple skeletons and make the decision of dose assignment. Simulation studies show that the proposed method yields higher MTD selection percentages and also assigns more patients to the MTD than the conventional dose-finding method, which does not borrow information across trials.

This article focuses on ethnic heterogeneity, but the proposed method can be used to handle other types of patient heterogeneity. For example, based on certain prognostic factors or biomarkers, patients often can be divided into several subgroups that have different levels of sensitivity to a drug. In this case, we can first use the conventional method to find the MTD in one subgroup, and then employ the proposed B-CRM to find the MTD in other subgroups. In some situations, certain modifications are needed for the proposed method to accommodate different types of prior information. For example, suppose a drug has been tested in adults, but we are interested in finding the MTD of that drug for children. Because children are typically more susceptible to toxicity and the MTD for children is most likely lower than that for adults, we may want to modify our three elicited skeletons such that the MTDs of the three skeletons are the same as, one level lower and two levels lower than the MTD of the landmark population. That is, we replace the skeleton that is one

level higher with a skeleton that is two levels lower than the MTD of the landmark population.

In the proposed B-CRM, we use the point estimate (i.e., posterior mean) of toxicity probability to determine the dose escalation and deescalation (i.e., step 2 of the dose-finding algorithm described in Section 2.4). Ishizuka and Ohashi [45] and Neuenschwander et al. [68] pointed out that the use of the point estimate may cause the tendency of aggressively allocating patients to toxic doses, at least under the logistic dose-toxicity model. To address this issue, Neuenschwander et al. [68] proposed to divide the toxicity probability into four intervals (i.e., under-dosing, targeted toxicity, excessive toxicity and unacceptable toxicity) and use the posterior probabilities of these intervals for the decision making in the CRM. The same strategy can be readily adopted here to enhance the performance of the B-CRM.

Appendix: Proof of Theorem 1

Let π_j denote the true toxicity probability of dose j. Because the nonparametric estimate is generally consistent (Bhattacharya and Kong, 2007), i.e., $\hat{\pi}_j^{(NP)} \to \pi_j$, the consistency of $\hat{\pi}$ depends on the property of the parametric estimate $\hat{\pi}_j^{(P)}$. When the probit model is correctly specified, $\hat{\pi}_j^{(P)}$ is a consistent estimate of π_j. Therefore, $\hat{\pi}_j$ is consistent since

$$\hat{\pi}_j = w_i \hat{\pi}_j^{(P)} + (1 - w_i)\hat{\pi}_j^{(NP)} \to w_i \pi_j + (1 - w_j)\pi_j = \pi_j.$$

We now show that $\hat{\pi}_j$ is still consistent when the probit model is misspecified. In this case, the parametric estimate $\hat{\pi}_j^{(P)}$ is generally not consistent with $\hat{\pi}_j^{(P)} \to \pi_j^*$, where π_j^* is a constant not equal to π_j. Define y_{ij} as the binary toxicity indicator for

the ith subject treated at dose j. The likelihood ratio of the two binomial distributions, as shown in (3.4), can be rewritten using the Bernoulli density, as follows,

$$\lambda_j = \frac{\prod_{i=1}^{m_j} f(y_{ij}|\hat{\pi}_j^{(P)})}{\prod_{i=1}^{m_j} f(y_{ij}|\hat{\pi}_j^{(NP)})},$$

where $f(y_{ij}|\hat{\pi}_j^{(P)}) = (\hat{\pi}_j^{(P)})^{y_{ij}}(1-\hat{\pi}_j^{(P)})^{1-y_{ij}}$ and $f(y_{ij}|\hat{\pi}_j^{(NP)}) = (\hat{\pi}_j^{(NP)})^{y_{ij}}(1-\hat{\pi}_j^{(NP)})^{1-y_{ij}}$. Thus,

$$\log(\lambda_j) = \sum_{i=1}^{m_j} \log \frac{f(y_{ij}|\hat{\pi}_j^{(P)})}{f(y_{ij}|\hat{\pi}_j^{(NP)})},$$

where each term in the summation has a mean of

$$\begin{aligned} E\left(\log\left(\frac{f(y_{ij}|\hat{\pi}_j^{(P)})}{f(y_{ij}|\hat{\pi}_j^{(NP)})}\right)\right) &\rightarrow E\left(\log\left(\frac{f(y_{ij}|\pi_j^*)}{f(y_{ij}|\pi_j)}\right)\right) \\ &= E\left(\log\left(\frac{f(y_{ij})}{f(y_{ij}|\pi_j)}\right) - \log\left(\frac{f(y_{ij})}{f(y_{ij}|\pi_j^*)}\right)\right) \\ &= H(\pi_j) - H(\pi_j^*), \end{aligned}$$

with $H(\theta)$ denoting the Kullback-Leibler information of the form

$$\begin{aligned} H(\theta) &= E\left(\log\left(\frac{f(y_{ij})}{f(y_{ij}|\theta)}\right)\right) \\ &= \int \log\left(\frac{f(y_{ij})}{f(y_{ij}|\theta)}\right) f(y_{ij})\,\mathrm{d}y_{ij} \end{aligned}$$

Based on Jensen's inequality, $H(\theta)$ is minimized at the true parameter value, $\theta = \pi_j$, with a minimum value of 0. Therefore,

$$E\left(\log\left(\frac{f(y_{ij}|\hat{\pi}_j^{(P)})}{f(y_{ij}|\hat{\pi}_j^{(NP)})}\right)\right) < 0.$$

That is, $\log(\lambda_j)$ is the sum of m_j iid random variables with negative mean. By the law of large numbers, $\log(\lambda_j) \to -\infty$ as $m_j \to \infty$. As a result, $w_j = \lambda_j/(1+\lambda_j) \to 0$. So we have

$$\hat{\pi}_j = w_j \hat{\pi}_j^{(P)} + (1 - w_j)\hat{\pi}_j^{(NP)} \to 0\pi_j^* + (1-0)\pi_j = \pi_j.$$

Table 3.1: Simulation study comparing the CRM, CRM using an informative prior (IP-CRM) and B-CRM. The underlined dose is the target dose.

			Dose level					
Scenario	Design		1	2	3	4	5	6
1		Pr(toxicity)	0.04	0.08	0.15	<u>0.33</u>	0.45	0.60
	CRM	selection(%)	0.0	2.3	23.0	**48.3**	22.6	3.4
		# of patients	0.1	2.0	7.1	7.0	3.9	0.9
	IP-CRM	selection(%)	0.0	0.3	25.3	**72.4**	2.0	0.0
		# of patients	0.0	0.4	7.9	12.4	0.4	0.0
	B-CRM	selection(%)	0.0	1.0	26.5	**56.1**	15.6	0.8
		# of patients	0.1	0.5	8.4	9.1	2.6	0.3
2		Pr(toxicity)	0.02	0.05	0.08	0.10	<u>0.30</u>	0.45
	CRM	selection(%)	0.0	0.1	1.4	15.5	**53.7**	29.2
		# of patients	0.0	0.8	4.1	4.5	6.5	5.0
	IP-CRM	selection(%)	0.0	0.0	3.3	59.9	**36.8**	0.0
		# of patients	0.0	0.1	4.4	12.2	4.3	0.0
	B-CRM	selection(%)	0.0	0.0	2.3	21.1	**59.5**	17.1
		# of patients	0.0	0.1	4.5	6.1	7.6	2.6
3		Pr(toxicity)	0.05	0.12	<u>0.25</u>	0.42	0.55	0.65
	CRM	selection(%)	0.6	11.4	**48.0**	31.9	6.9	0.2
		# of patients	0.4	4.0	8.9	5.2	2.1	0.3
	IP-CRM	selection(%)	0.1	6.1	**54.0**	39.1	0.7	0.0
		# of patients	0.0	1.8	11.3	7.7	0.1	0.0
	B-CRM	selection(%)	0.1	8.7	**56.3**	31.7	2.9	0.1
		# of patients	0.3	1.9	11.6	6.1	1.0	0.0
4		Pr(toxicity)	0.02	0.03	0.04	0.06	0.10	<u>0.33</u>
	CRM	selection(%)	0.0	0.0	0.1	1.1	18.8	**80.0**
		# of patients	0.0	0.4	3.5	3.2	4.5	9.3
	IP-CRM	selection(%)	0.0	0.0	0.4	40.2	59.4	**0.0**
		# of patients	0.0	0.0	3.5	10.4	7.1	0.0
	B-CRM	selection(%)	0.0	0.0	0.2	3.0	33.7	**63.1**
		# of patients	0.0	0.0	3.5	4.2	6.0	7.3
5		Pr(toxicity)	0.15	<u>0.26</u>	0.50	0.60	0.70	0.75
	CRM	selection(%)	16.8	**45.6**	21.4	1.7	0.2	0.0
		# of patients	3.3	7.8	6.4	0.9	0.2	0.0
	IP-CRM	selection(%)	9.7	**56.9**	31.1	1.9	0.0	0.0
		# of patients	1.7	8.4	9.4	1.4	0.0	0.0
	B-CRM	selection(%)	15.9	**58.6**	21.6	0.6	0.1	0.0
		# of patients	3.2	7.8	8.7	0.9	0.0	0.0
6		Pr(toxicity)	<u>0.30</u>	0.46	0.55	0.65	0.75	0.85
	CRM	selection(%)	**40.9**	19.0	3.5	0.3	0.0	0.0
		# of patients	6.3	5.4	4.1	0.5	0.1	0.0
	IP-CRM	selection(%)	**48.2**	35.0	9.2	0.8	0.0	0.0
		# of patients	5.7	7.6	6.7	0.7	0.0	0.0
	B-CRM	selection(%)	**48.6**	21.5	5.2	0.2	0.0	0.0
		# of patients	6.4	5.3	6.0	0.5	0.0	0.0

Table 3.2: Sensitivity analysis of the CRM, IP-CRM and B-CRM, given landmark trial data $(m_1, \ldots, m_6) = (3, 3, 6, 3, 0, 0)$ and $(x_1, \ldots, x_6) = (0, 0, 1, 2, 0, 0)$. The underlined dose is the target dose.

			Dose level					
Scenario	Design		1	2	3	4	5	6
1		Pr(toxicity)	0.05	0.12	<u>0.35</u>	0.42	0.55	0.65
	CRM	selection(%)	0.4	23.6	**46.9**	23.6	5.1	0.1
		# of patients	1.4	6.8	7.1	4.3	1.1	0.2
	IP-CRM	selection(%)	0.0	21.5	**76.2**	2.3	0.0	0.0
		# of patients	0.2	7.2	13.2	0.4	0.0	0.0
	B-CRM	selection(%)	0.0	22.3	**59.6**	16.5	1.4	0.0
		# of patients	0.4	7.7	9.6	2.9	0.3	0.0
2		Pr(toxicity)	0.04	0.08	0.15	<u>0.26</u>	0.45	0.60
	CRM	selection(%)	0.0	1.4	15.6	**47.9**	31.2	3.8
		# of patients	1.0	4.3	4.7	6.2	3.9	0.8
	IP-CRM	selection(%)	0.0	3.8	70.1	**25.3**	0.8	0.0
		# of patients	0.1	4.5	13.3	3.1	0.0	0.0
	B-CRM	selection(%)	0.0	3.0	23.7	**56.4**	16.0	0.9
		# of patients	0.2	4.7	6.6	7.1	2.2	0.2
3		Pr(toxicity)	0.15	<u>0.26</u>	0.50	0.60	0.70	0.75
	CRM	selection(%)	19.6	**50.5**	23.0	2.1	0.2	0.0
		# of patients	5.1	9.1	4.8	1.3	0.1	0.0
	IP-CRM	selection(%)	9.1	**63.4**	26.8	0.0	0.0	0.0
		# of patients	2.1	12.4	6.4	0.0	0.0	0.0
	B-CRM	selection(%)	10.0	**64.3**	22.4	1.1	0.0	0.0
		# of patients	2.5	12.3	5.2	0.5	0.0	0.0
4		Pr(toxicity)	0.02	0.05	0.08	0.10	<u>0.30</u>	0.45
	CRM	selection(%)	0.0	0.1	1.2	14.9	**50.3**	33.5
		# of patients	0.5	3.6	3.3	4.3	5.6	3.7
	IP-CRM	selection(%)	0.0	0.8	49.0	34.0	**16.2**	0.0
		# of patients	0.0	3.7	11.3	5.4	0.6	0.0
	B-CRM	selection(%)	0.0	0.4	5.6	33.3	**46.7**	14.0
		# of patients	0.1	3.7	4.7	5.7	5.4	1.4
5		Pr(toxicity)	<u>0.30</u>	0.46	0.55	0.65	0.75	0.85
	CRM	selection(%)	**48.1**	17.7	1.9	0.2	0.1	0.0
		# of patients	9.4	6.0	1.1	0.3	0.0	0.0
	IP-CRM	selection(%)	**52.9**	32.7	4.6	0.0	0.0	0.0
		# of patients	8.7	9.5	1.8	0.0	0.0	0.0
	B-CRM	selection(%)	**54.4**	22.3	3.1	0.3	0.0	0.0
		# of patients	8.9	8.0	1.5	0.1	0.0	0.0
6		Pr(toxicity)	0.02	0.03	0.04	0.06	0.10	<u>0.25</u>
	CRM	selection(%)	0.0	0.0	0.2	1.6	14.9	**83.3**
		# of patients	0.3	3.3	3.1	3.3	3.9	7.1
	IP-CRM	selection(%)	0.0	0.1	29.6	30.7	39.6	**0.0**
		# of patients	0.0	3.3	9.2	7.2	1.2	0.0
	B-CRM	selection(%)	0.0	0.0	1.0	14.3	22.7	**62.0**
		# of patients	0.0	3.4	3.8	4.6	4.5	4.7

Table 3.3: The number of DLTs at six doses in the phase I trial of BKM120 for patients with advanced solid tumors.

	Dose (mg)					
	12.5	25	50	80	100	150
Number of patients	1	2	5	6	17	4
Number of DLTs	0	0	0	1	4	2

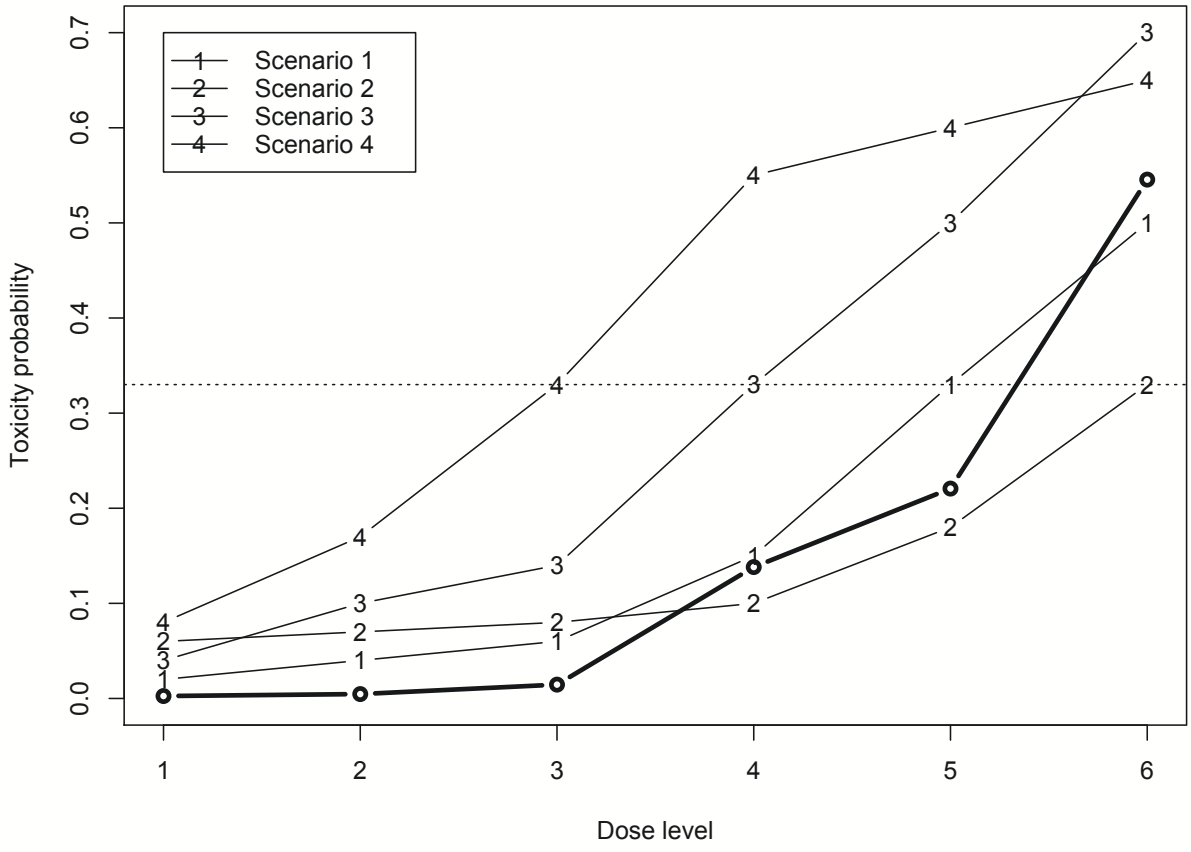


Figure 3.1.: Four dose-toxicity curves for the new population and the estimate of the dose-toxicity curve for the landmark population (represented by the thick line). The horizontal dotted line indicates the target toxicity probability.

4. A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials

4.1 Introduction

This chapter is based on "A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials" published in Journal of Royal Statistical Science- Series C (2017) [82] coauthored with Ying Yuan and Jielai Xia. Permission from the journal has been granted for use in conjunction with the thesis.

As introduced in Chapter 1, biosimilar is a term that describes the equivalence of a generic version to an innovator's biologic drug product; biosimilars are close, but not exact copies of biologic drugs already on the market. Since the patents of many blockbuster proprietary biologic products reach their expirations, biosimilars provide great potential to increase the accessibility of biologic products for patients with life-threatening diseases. A biosimilar must demonstrate "biosimilarity" to its innovative reference product in terms of quality characteristics, biological activity, safety and efficacy based on comprehensive comparability studies before going to the market. However, the development of biological products is much more complicated than that of conventional small-molecule-based drugs, and biologics are sensitive to small procedural or environmental changes during the manufacturing process, the conventional approach to evaluating bioequivalence based on pharmacokinetic responses cannot be directly applied to establish biosimilarity. Furthermore, guidelines for statistical methods to establish biosimilarity remain nonspecific because of the newness

of biosimilars, even though regulatory agencies, such as the U.S. Food and Drug Administration (FDA), the European Medicines Agency, and the World Health Organization, have provided detailed guidance for demonstrating comparability in terms of quality, safety and efficacy. It is therefore of high urgency to develop appropriate and reliable statistical methodologies for developing biosimilars.

In this project, we propose a two-arm randomized Bayesian group sequential design to evaluate the biosimilarity between an investigational biosimilar and the innovative reference drug. Biosimilar trials come with several challenges that are beyond the scope of the conventional randomized comparative trial design. First, when a biosimilar is ready to be tested in a randomized trial, the innovative reference drug has been in the market for many years and a huge amount of data on that drug has accumulated. It is critical to incorporate these rich historical data into the biosimilar trial design to improve trial efficiency. An efficient trial design not only leads to tremendous cost saving for the pharmaceutical industry, but translates into saving lives because it allows patients to access the biosimilars earlier by expediting their development. Another challenge when designing biosimilar trials is determining how to quantify and monitor the biosimilar during the trial.

The remainder of this chapter is organized as follows. In Section 4.2, we briefly review the power prior and propose the CPP. In Section 4.3, we propose the BBI for assessing the similarity between the biosimilar and the innovative reference drug, based on which we develop a Bayesian adaptive design for two-arm randomized biosimilar trials. We investigate the operating characteristics of the proposed design using simulation studies in Section 4.4. In Section 4.5, we apply the proposed methodology to design a biosimilar trial for treating arthritis, and conclude with a brief discussion in Section 4.6.

4.2 Methods

4.2.1 Power prior

A power prior provides an intuitive approach for borrowing information from historical data. Let θ denote the parameter of interest, and $\pi_0(\theta)$ denote the prior distribution of θ (before accounting for the historical data), which is typically specified as the noninformative or flat prior. Let D_0 denote the historical data and D denote the data from the current trial. The basic idea of the power prior is straightforward: update $\pi_0(\theta)$ using D_0, and then use the resulting posterior as the (power) prior to make posterior inference so that the information of D_0 is incorporated into the analysis of D. More precisely, the power prior can be written as

$$\pi(\theta|D_0, \delta) \propto L(\theta|D_0)^{\delta}\pi_0(\theta) \tag{4.1}$$

where $L(\theta|D_0)$ is the likelihood of θ conditional on historical data D_0, and $\delta \in [0, 1]$ is the power parameter, which controls how much information we borrow from D_0. When $\delta = 1$, we fully borrow information from D_0 and when $\delta = 0$, we do not borrow any information from D_0. When D_0 come from the exponential distribution family, e.g., a normal or binomial distribution, δ can be interpreted as the fraction of the information borrowed from D_0. For example, for n normally distributed observations with mean θ, $L(\theta|D_0)^{\delta}$ is equivalent to a likelihood obtained by inflating the variance with a factor of $1/\delta$, or, equivalently, a discounted historical sample size of $n\delta$.

As the value of δ is typically unknown in practice, the fully Bayesian approach treats δ as an unknown parameter [72,73] and assigns it a prior distribution $\pi(\delta)$, e.g., $\pi(\delta) \sim Unif(0, 1)$, yielding the power prior as follows:

$$\pi(\theta, \delta|D_0) = \frac{L(\theta|D_0)^{\delta}\pi_0(\theta)\pi(\delta)}{C(\delta)}, \tag{4.2}$$

where $C(\delta) = \int_\Theta L(\theta|D_0)^\delta \pi_0(\theta) d\theta$ is a normalizing constant. Ye et al. [67], Duan et al. [69][70] and Neuenschwander and Spiegelhalter [68] noted that it is critical to include the normalizing constant $C(\delta)$ in (4.2), and that ignoring $C(\delta)$ leads to pathological priors, such as in the early literature on power priors [72,73,74]. Given data D from the current trial, the posterior distribution of θ and δ is given by $\pi(\theta, \delta|D, D_0) \propto L(\theta|D)\pi(\theta, \delta|D_0)$.

Although the power prior is intuitive and conceptually attractive, using it in practice is tricky. Neuenschwander and Spiegelhalter [68] found that the power parameter δ cannot be estimated accurately based on D and D_0, even when the sample size of each data set is large. In other words, the power prior cannot appropriately determine how much information we should borrow from D_0. This led these authors to recommend fixing the value δ *a priori*, rather than estimating it from the data. Ideally, δ should be set close to 1 if D and D_0 are congruent, and close to 0 if they are not. Unfortunately, this is difficult to implement in practice because *a priori* we typically do not know the degree of congruence between D_0 and D. As a result, Neuenschwander and Spiegelhalter [68] concluded that "though the δ is easy to interpret, its elicitation is challenging."

4.2.2 Calibrated power prior

To address the aforementioned issues, we propose the CPP, for which δ is defined as a function of a congruence measure between D_0 and D. The key to our approach is that the function which links δ and the congruent measure is *prespecified* and calibrated by simulation such that when D_0 is congruent with D, the CPP strongly borrows information from D_0, thereby improving power, and when D_0 is not congruent with D, the CPP borrows little information from D_0, thereby controlling the type I error rate.

We first introduce a measure of congruence between $D_0 = (x_1, \cdots, x_m)$ and $D = (y_1, \cdots, y_n)$, where x and y can be continuous or binary variables. A natural measure of congruency between D_0 and D is the Kolmogorov-Smirnov (KS) statistic, a nonparametric statistic for testing whether D_0 and D have the same probability distribution. We note that the KS statistics is not the only choice, and other reasonable measure of congruency can also be used. Specifically, for a real number t, letting $F_m(t) = \sum I(x_j \leq t)/m$ and $G(t) = \sum I(y_i \leq t)/n$ denote the empirical distribution functions for D_0 and D, the KS statistic is defined as $S_{KS} = \max_{-\infty<t<\infty}\{|F(t) - G(t)|\}$. Letting $Z_{(1)} \leq \cdots \leq Z_{(N)}$ denote the $N = m + n$ ordered values for the combined sample of D_0 and D, the KS statistic can be calculated as

$$S_{KS} = \max_{i=1,\ldots,N}\{|F(Z_{(i)}) - G(Z_{(i)})|\}.$$

The KS statistic measures the discrepancy or incongruence between the distributions of D_0 and D. A large value of S_{KS} indicates a larger incongruence between the distributions of D_0 and D. In our approach, we adopt a scaled KS statistic, defined as

$$S = \max(m, n)^{\frac{1}{4}} S_{KS}.$$

The reason we choose to use S, rather than the original KS statistic, is to ensure that the resulting CPP has a desirable property when borrowing information from D_0, as described later in Theorem 1.

We link the power parameter δ with S through

$$\delta = g(S; \phi), \tag{4.3}$$

where $g(\cdot)$ is a monotonically increasing function with parameter ϕ, known as *calibration function*, satisfying the following requirements: when S is small, which indicates that D_0 and D are congruent, $g(S; \phi)$ is close to 1 to strongly borrow information

from D_0; and when S is large, which indicates that D and D_0 are incongruent, we require that $g(S;\phi)$ is close to 0 to acknowledge the difference between D and D_0 and refrain from borrowing information from D_0. Although many different forms of $g(\cdot)$ satisfy these requirements, one particular function form that is simple and yields good operating characteristics is the two-parameter reciprocal exponential model

$$\delta = g(S;\phi) = \frac{1}{1+\exp\{a+b\times\log(S)\}}, \tag{4.4}$$

where $\phi = (a,b)$ are tuning parameters that control the relationship between δ and S. We require $b > 0$ to ensure that the larger incongruence between D_0 and D leads to a smaller value of δ. The procedure to determine the values of a and b is describe later. The proposed CPP can be generally expressed as

$$\pi(\theta|D_0,a,b) = L(\theta|D_0)^{[1+\exp\{a+b\times\log(S)\}]^{-1}}\pi_0(\theta).$$

The CPP has the following large-sample property. The proof is provided in the Appendix.

Theorem 1 When D_0 and D have the same distribution (i.e., are congruent), δ in (4.4) converges to 1 and thus the CPP fully borrows information from D_0; and when D_0 and D have different distributions (i.e., are incongruent), δ converges to 0 and thus the CPP does not borrow any information from D_0.

In contrast, the original power prior may not have this desirable convergence property. This is because given only two data sets, the heterogeneity between the data sets, and thus δ, cannot be estimated precisely, even when the sample size of each data set is large. As noted by Neuenschwander, Branson, & Spiegelhalter [68], this is analogous to a hierarchical (random-effects meta-analytic) model, for which it is difficult to obtain a reasonably precise estimate for the between-trial variability if only a few trials are available.

The CPP follows the spirit of empirical Bayesian methodology in the sense that it depends on the observed data D through S. However, unlike the typical empirical Bayesian methodology, the determination of the tuning parameters a and b does not rely on the data D actually observed in the current study. We calibrate the value of a and b using simulated data, as follows. We first consider the case in which $D_0 = (x_1, \cdots, x_m)$ and $D = (y_1, \cdots, y_n)$ are normally distributed, with $x_i \sim N(\mu_0, \sigma_0^2)$ and $y_j \sim N(\mu_0 + \gamma, \sigma_0^2)$, $i = 1, \cdots, m$ and $j = 1, \cdots, n$. Given historical data D_0, the values of the tuning parameters a and b are calibrated as follows,

1. Estimate the mean and variance of D_0 by $\hat{\mu}_0 = \bar{x}$ and $\hat{\sigma}_0^2 = \sum_{i=1}^m (x_i - \bar{x})/(m-1)$ with $\bar{x} = \sum_{i=1}^m x_i/m$.

2. Elicit from subject experts the maximum practically negligible mean difference γ, denoted as γ_c, such that D and D_0 can be regarded as congruent, and the minimal value of γ, denoted as $\gamma_{\bar{c}}$, such that D is deemed to be substantially different (i.e., not congruent) from D_0. As we describe later, this elicitation procedure is simple to implement for biosimilar studies.

3. Generate R replicates of D by simulating $(y_1, \cdots, y_n)$ from $N(\hat{\mu}_0 + \gamma_c, \hat{\sigma}_0^2)$, and calculate the KS statistics between each of these R simulated dataset and D_0. Let $S^*(\gamma_c)$ denote the median of the R resulting KS statistics.

4. Repeat step 3 by replacing γ_c with $\gamma_{\bar{c}}$, and let $S^*(\gamma_{\bar{c}})$ denote the median of the R resulting KS statistics.

5. Solve a and b in (4) based on the following two equations:

$$\delta_c = g(S^*(\gamma_c); \phi) \tag{4.5}$$

$$\delta_{\bar{c}} = g(S^*(\gamma_{\bar{c}}); \phi), , \tag{4.6}$$

where δ_c is a large constant close to 1 (e.g., 0.98), and $\delta_{\bar{c}}$ is a small constant close to 0 (e.g., 0.01). The rationale is that when D_0 and D are congruent (i.e., $\gamma = \gamma_c$), we want to strongly borrow information from D_0 (i.e., δ is close to 1), and when D_0 and D are not congruent (i.e., $\gamma = \gamma_{\bar{c}}$), we want to refrain from borrowing information from D_0 (i.e., δ is close to 0) to avoid bias and inflate the type I error rate. Solving (4.5) and (4.6) leads to the values of a and b as follows,

$$a = \log(\frac{1-\delta_c}{\delta_c}) - \frac{\log(\frac{(1-\delta_c)\delta_{\bar{c}}}{(1-\delta_{\bar{c}})\delta_c})\log(S^*(r_c))}{\log(\frac{S^*(r_c)}{S^*(r_{\bar{c}})})} \tag{4.7}$$

$$b = \frac{\log(\frac{(1-\delta_c)\delta_{\bar{c}}}{(1-\delta_{\bar{c}})\delta_c})}{\log(\frac{S^*(r_c)}{S^*(r_{\bar{c}})})}. \tag{4.8}$$

Several remarks are warranted. First, we can see that the calibration of a and b do not depend on D, the data collected from the current study. This is an important and very desirable property because it allows the investigator to determine the values of a and b and to include them in the study protocol before the onset of the study. This will address the major concern about the methods of borrowing information from historical data, that is, the method could be abused by choosing the degree of borrowing to favor a certain result, e.g., statistically significant results. Second, in step 2, γ_c and $\gamma_{\bar{c}}$ are similar to the effect sizes (i.e., mean differences) that are routinely used in power calculations, and thus can be readily elicited from subject experts. This elicitation is particularly straightforward for biosimilar studies because in order to assess biosimilarity, *a priori*, it is imperative to specify the biosimilar margin/criterion (i.e., define the level of similarity required). In practice, we often use the 0.80/1.25 rule, that is, the investigational biosimilar is regarded as being similar to the reference agent if the difference between their (log-transformed) means

is within $(\log(0.80) = -0.223, \log(1.25) = 0.223)$. In this case, it is natural to choose $\gamma_{\bar{c}} = 0.223$, and set γ_c at the value that represents a practically negligible difference.

In the proposed procedure, a and b are solved on the basis of two elicited values of γ (i.e., γ_c and $\gamma_{\bar{c}}$). If desirable, more than two values of γ can be elicited and paired with desirable degrees of information borrowing from D_0, for example, $\gamma = (0.223, 0.2, 0.15, 0.1, 0)$ and $\delta = (0, 0.25, 0.50, 0.75, 1)$. This will result in more than two equations of the form of (4.5) and (4.6). In this case, the least squares method can be used to solve a and b. We note that a common variance is assumed to simulate D in step 3. That assumption can be easily relaxed by using a different value of variance for simulating D. However, as the goal of the above procedure is to calibrate the value of a and b, not to assess a biosimilar, the common variance assumption is not critical, as shown later in the sensitivity analysis.

The above calibration procedure can also be used to handle the case in which D and D_0 are binary endpoints with minor modifications. Details are provided in the Appendix. In the next section, we describe how to use the CPP to design two-arm randomized biosimilar trials.

4.3 Bayesian design for comparative biosimilar trials

Consider a biosimilar trial in which patients are randomized to receive an investigational biosimilar (T) or an innovative reference (R) drug. Let Y_T and Y_R denote the primary clinical efficacy endpoints for T and R, respectively, which can be a continuous or binary variable. Denote $\mu_k = E(Y_k)$ for $k = T, R$. We assume that historical data $D_0 = (x_1, \cdots, x_m)$ are available for R.

Before describing our design, we propose a new measure, the *Bayesian Biosimilarity Index* (BBI), to quantify the similarity between T and R,

$$\text{BBI} = \Pr(\lambda_L < \mu_T/\mu_R < \lambda_U | data),$$

where λ_L and λ_U are the prespecified biosimilarity limits. In practice, (λ_L, λ_U) are often chosen as (80%, 125%). For log-transformed normal data, the BBI can be equivalently defined as BBI $= Pr(\lambda_L^* < \mu_T - \mu_R < \lambda_U^* | data)$, where biosimilarity limits $\lambda_L^* = \log(\lambda_L)$ and $\lambda_U^* = \log(\lambda_U)$ and are often chosen as (-0.223, 0.223). Compared to the existing approaches based on the frequentist confidence interval of μ_T/μ_R, one important advantage of the BBI is its intuitive interpretation and its ability to define and assess biosimilarity using easy-to-understand probability statements. Specifically, the BBI represents the probability that T and R are biosimilar (i.e., located within the prespecified biosimilarity limits), given the observed data. For example, BBI=95% means that there is 95% chance that R and T are similar based on the observed data. In contrast, the 95% confidence interval of μ_T/μ_R only tell us the range of the values that have 95% chance of covering the true value of μ_T/μ_R under repeated sampling. It does not tell us how likely it is that μ_T/μ_R is located within the prespecified biosimilarity limits (i.e., satisfies the biosimilar criterion). We may compare the confidence interval with (λ_L, λ_U) to see whether the former is located within the latter, but the confidence interval still does not tell us the probability that R and T are similar.

With the BBI in hand, the proposed Bayesian design is described as follows, assuming that K interim looks are planned for the trial after $n_1, \cdots, n_K$ patients have been enrolled into arms T and R.

1. Enroll $2n_1$ patients and randomize them to T and R arms.

2. Given the kth interim data $D_T(n_k) = (y_{T,1}, \cdots y_{T,n_k})$ $D_R(n_k) = (y_{R,1}, \cdots y_{R,n_k})$, $k = 1, \cdots, K$

    i. (*Futility stopping*) If BBI $< C_f$, terminate the trial early and conclude that T is not similar to R, where C_f is a probability cutoff for futility stopping,;

ii. (*Superiority stopping*) If BBI $> C_s$, terminate the trial early and conclude that T and R are similar, where C_s is a probability cutoff for superiority stopping;

iii. Otherwise, continue to enroll patients until the next interim analysis is reached.

3 Once the maximum sample size is reached, compute the BBI based on all observed data. If BBI $> C_s$, conclude that T and R are similar; otherwise, they are not similar.

To ensure that the design possesses good frequentist operating characteristics, probability cutoffs C_f and C_s should be calibrated through simulations to achieve desirable type I and II error rates. This simulation-based calibrated procedure is widely used in Bayesian clinical trial designs [76, 77]. The software to implement the proposed Bayesian biosimilar design (written in R) will be available at http://odin.mdacc.tmc.edu/∼yyuan/. The details of calculating BBI at each interim are provided in Appendix.

4.4 Simulation studies

4.4.1 Simulation setting

We investigated the operating characteristics of the proposed Bayesian design via simulation studies. We considered both the normally distributed endpoint and the binary endpoint. For the normally distributed endpoint, the maximum total sample size was 240, with two interim analyses conducted when 80 and 160 patients were enrolled. Patients were equally randomized into arms T and R. We generated Y_R from $N(\mu_R, 0.5^2)$, with $\mu_R = 0$, and generated Y_T from $N(\mu_T, 0.5^2)$, with $\mu_T = -0.223$, -0.115, 0, 0.115, and 0.223. We adopted the 0.80/1.25 rule to define biosimilarity such that T and R are similar if $-0.223 < \mu_R - \mu_T < 0.223$, assuming that Y_T and Y_R are

log-transformed data. In other words, T and R are similar when $\mu_T = -0.115$, 0 and 0.115, and not similar when $\mu_T = -0.223$ and 0.223. We generated historical data X from $N(\mu_0, 0.5^2)$ with $\mu_0 = 0$, -0.5, -0.3, 0.3, 0.5 and sample size N_0 =300 and 500. To obtain the CPP, we elicited $\gamma_c = 0$ and $\gamma_{\bar{c}} = 0.223$ with $\delta_c = 0.99$ and $\delta_{\bar{c}} = 0.001$. The resulting tuning parameters a and b are displayed in Table 4.1. In our Bayesian design, we set $C_f = 0.4$ and $C_s = 0.955$, which are chosen by calibrating the type I error rate to the nominal value of 5% when $\mu_T = -0.223$ and 0.223.

For the binary endpoint, the maximum total sample size was 1800, with two interim analyses conducted when 600 and 1200 patients were enrolled. To obtain reasonable power, such as 80%, the binary endpoint requires a much larger sample size than the normal endpoint. We generated Y_R from the Bernoulli distribution $Ber(\mu_R)$, with $\mu_R = 0.5$, and generated Y_T from $Ber(\mu_T)$, with $\mu_T = 0.4$, 0.45, 0.5, 0.565, and 0.625. Under the 0.80/1.25 rule, T and R are similar when $\mu_T = 0.45$, 0.5 and 0.565 because in these cases, $0.8 < \mu_T/\mu_R < 1.25$, and are not similar when $\mu_T = 0.4$ and 0.625 because $\mu_T/\mu_R \leq 0.8$ or ≥ 1.25. We generated historical data X from $Ber(\mu_0)$, with $\mu_0 = 0.5$, 0.2, 0.8, 0.1 and 0.9 and sample size $m = 600$ and 1000. To obtain the CPP, we elicited $\gamma_c = 0$ and $\gamma_{\bar{c}} = 0.223$, with $\delta_c = 0.99$ and $\delta_{\bar{c}} = 0.001$. We set $C_f = 0.8$ and $C_s = 0.96$, to ensure appropriate type I error rates.

We compared the proposed CPP design with two alternative designs. The first alternative design is called the no borrowing (NB) design, which is the same as the proposed design except that it ignores historical data. The second design uses the standard power prior (denoted as the PP design) to borrow information from the historical data. The PP design is a fully Bayesian approach, under which δ is treated as an unknown parameter and assigned with a uniform prior $\delta \sim Unif(0, 1)$.

4.4.2 Simulation results

Table 4.2 shows the results for normal endpoints based on 10,000 simulated trials. As the NB design is not affected by the historical data, its results are shown only once at the top of the table. In scenario 1, the historical data D_0 are congruent with the reference arm data D_R (i.e., $\mu_0 = 0 = \mu_R$). The proposed CPP design had higher power to detect the similarity between R and T than the NB design. Specifically, when the R and T are similar (i.e., μ_T = 0, 0.115 and -0.115) and the sample size of the historical data $N_0 = 300$, the powers of the CPP design were 67.5%, 96.9% and 67.1%, respectively, while those of the NB design were 58.2%, 95.5% and 58.9%. The gain was more obvious when $N_0 = 500$, under which the powers of the CPP design were improved to 69.0%, 97.6% and 69.2%. Such a power improvement is impressive given that the CPP design used smaller sample sizes than the NB design. For example, when $N_0 = 300$, the sample sizes of the CPP designs were 88.79, 76.51 and 88.54 when $\mu_T = -0.115$, 0 and 0.115, while those of the NB design were 93.32, 85.31 and 93.26. When R and T are not similar (i.e., $\mu_T = -0.223$ or 0.223), both the NB and CPP designs controlled the type I error rate (i.e., concluding that T and R are similar when they are actually not) close to the nominal value of 5%. In scenario 1 (i.e., D_0 and D_R are congruent), the PP design yielded higher power than the CPP and NB designs and an appropriate type I error rate. However, when D_0 and D_R are not congruent, the PP design led to a substantially inflated type I error rate. For example, in scenario 2, when D_0 and D_R are not congruent, with $\mu_0 = -0.5$ (recall that $\mu_R = 0$), the type I error rate of the PP design was 17.4% and 22.0% when $N_0 = 300$ and 500. This result confirms the previous finding regarding the standard power prior: the power parameter cannot be precisely estimated based on the data, and thus cannot appropriately determine how much information should be borrowed from D_0. In contrast, the proposed design correctly recognized that D_0 and D are not congruent and thus no information should be borrowed. This is reflected

by the appropriate type I error rate of the CPP design (i.e., 5.4% and 5.4%) when $\mu_T = 0.223$ and -0.223. The power of the CPP design is comparable to that of the NB design when R and T are similar (i.e., $\mu_T = -0.115$, 0 or 0.115). In scenarios 3 to 5, D_0 and D_R are not congruent, with different values of μ_0. We observed similar results. That is, the CPP design well controlled the type I error rate and yielded power comparable to that of the NB design. The PP design had high power, but did not control the type I error rate. Figure 4.1 shows the power curve of the CPP design under different values of μ_0, with the NB design as the reference. We can see that the CPP yielded higher power than the NB design when D_0 and D are congruent (i.e., $\mu_0 = 0$), with well controlled type I error rates.

Table 4.3 provides the simulation results for binary endpoints. The results are generally similar to those for the normal endpoint. For example, in scenario 1, the historical data D_0 are congruent to the reference arm data D_R (i.e., $\mu_0 = 0.5 = \mu_R$). Given the sample size of the historical data $N_0 = 1000$, the powers of the proposed CPP design were 3.3% to 17.8% higher than those of the NB design when R and T are similar (i.e., $\mu_T = 0$, 0.115 and -0.115). In addition, the CPP design controlled the type I error rate close to 5% when R and T are not similar (i.e., $\mu_T = 0.4$ or 0.625). Again, although the PP design yielded higher statistical power when D_0 and D_R are congruent (i.e., scenario 1), it led to dramatically inflated type I error rates (see scenarios 2-5). For example, in scenario 2, with $N_0 = 1000$, the type I error rate of the PP design was 15.2 when $\mu_T = 0.4$. Figure 4.2 shows the power curves of the CPP design under different values of μ_0, with the NB design as the reference.

4.4.3 Sensitivity analysis

For the normal endpoint, our calibration procedure (Section 4.2.2) for the CPP assumes that D has the same variance as D_0. We conducted a sensitivity analysis to evaluate the performance of the proposed design when D and D_0 actually have

different variances. Simulation results (see Table 4) show that our design controlled the type I error rate at the nominal level of 5% when R and T are not similar, and yielded reasonable power when R and T are similar. In contrast, the PP approach led to inflated type I error rates up to 15%.

We also conducted a sensitivity analysis to evaluate the impact of the specification of δ_c and $\delta_{\bar{c}}$ (in step 5 of the CPP procedure in Section 2.2) on the performance of the design. We considered three different specifications of δ_c and $\delta_{\bar{c}}$, i.e., $(\delta_c, \delta_{\bar{c}}) = (0.99, 0.001)$, (0.95, 0.005), (0.95, 0.0001). Figure 4.3 shows that the operating characteristics of the design are very similar under different values of δ_c and $\delta_{\bar{c}}$, suggesting that our design is robust to the specification of δ_c and $\delta_{\bar{c}}$.

4.5 Application

Adalimumab(Humira) is the first fully human monoclonal antibody drug approved by the FDA in 2002 for treating rheumatoid arthritis (RA) and other types of arthritis. RA is an autoimmune disease characterized by progressive inflammatory synovitis of the joints that may result in erosion of articular cartilage and subchondral bone. RA is a relatively common disease with prevalence from 0.4% to 1.3% worldwide, and more than 200,000 cases per year in the United States [78]. Due to the high cost of Humira, e.g., approximately $3,000 per month in 2015, a substantial portion of patients cannot receive this effective treatment, especially in developing countries such as China. Given that the patent for this antibody expires in 2016, our collaborators in China are interested in developing a biosimilar monoclonal antibody of Humira to reduce the cost of the drug and allow more patients to benefit from the treatment.

A two-arm randomized clinical trial was proposed to evaluate the biosimilarity between the test agent and Humira. The primary endpoint is clinical response at week 24, a binary outcome indicating whether the patient achieves an improvement of at least 20% in the American College of Rheumatology core criteria (ACR20) from

baseline to week 24. Patients who did not achieve an ACR20 response, who withdrew from the study, or who received "rescue treatment with traditional disease-modifying antirheumatic drug therapy on or after week 16 were classified as nonresponders. A maximum of 345 patients will be equally randomized to receive the test agent or adalimumab administered at 20mg weekly. The available historical data were obtained from a randomized clinical trial and included information on 212 patients who were treated with adalimumab [79]. The response rate of ACR20 was 60.8% in the historical data. We applied the proposed methodology to design the trial. We determined the calibration function (4.4) using the procedure described in Section 2.2 and the Appendix. Based on the 0.80/1.25 rule, we set $\gamma_c = 0.99$ and $\gamma_{\bar{c}} = 0.001$, resulting in the solution $\hat{a} = 18.63$ and $\hat{b} = 5.53$. That is, the power parameter used in the trial is given by

$$\delta = \frac{1}{1 + \exp\{18.63 + 5.53 \times \log(S)\}}.$$

We examined the operating characteristics of the resulting CPP design under three scenarios (see Table 5), contrasted with the conventional no-borrowing (NB) design that ignores the historical data. In scenario 1, for which the test agent is biosimilar to adalimumab and the historical data is congruent with the control data (i.e., Humira arm), the proposed design yielded 81% power, whereas the NB design yielded 67% power, demonstrating that the use of historical data can substantially improve the power of the study. In scenario 2, the test agent is also biosimilar to adalimumab, but the historical data are not congruent to the control. The CPP and NB design yields similar power. Scenario 3 considers the case in which the historical data are congruent with the control data, but the test agent and adalimumab are not biosimilar. The CPP design well controlled the type I error rate below the nominal value of 5%, demonstrating that the CPP design correctly recognized that, in this case, no

information should be borrowed from the historical data to maintain an appropriate type I error rate.

4.6 Summary

We have proposed a Bayesian group sequential adaptive design for biosimilar trials. To incorporate rich historical data that are almost always available for biosimilar trials, we developed the CPP, which allows the design to adaptively borrow information from historical data. When the historical data are congruent with the new data collected from the trial, the CPP borrows information from the historical data and thus improves the power of the design; and when the historical data are not congruent with the new data from the trial, the CPP well controls the type I error rate. To facilitate trial monitoring, we proposed the BBI to measure the similarity between the biosimilar and the innovative reference drug. Our design evaluates the BBI in a group sequential fashion based on the accumulating interim data, and stops the trial early once there is enough information to conclude or reject the similarity. Our simulation studies show that the proposed design has desirable operating characteristics.

This article focuses on biosimilar trials. The proposed CPP approach can be used to adaptively borrow information from historical data in other settings. For example, in bridging clinical trials, as the landmark trial has been completed, we could use the CPP to design a follow-up trial (i.e., a bridging trial). We have considered binary and normal endpoints. The proposed approach can be extended to time-to-event endpoints as well. This will be the topic of our future research.

Appendix in Chapter 4

1. Proof of Theorem 1

Proof Supposing that $m, n \to \infty$ and $m/n \to O(1)$, that is, the sample sizes of D_0 and D increase on the same order. Without loss of generality, we assume that $m \geq n$. Thus,

$$\begin{aligned} S &= \max(m,n)^{1/4} S_{KS} && (4.9) \\ &= m^{1/4} S_{KS} && (4.10) \\ &= m^{-1/4}(m/n+1)^{1/2}\sqrt{\frac{mn}{m+n}} S_{KS} && (4.11) \end{aligned}$$

Smirnov (1939) showed that when D_0 and D have the same distribution (i.e., are congruent), $\sqrt{\frac{mn}{m+n}} S_{KS}$ converges in distribution to Kolmogorov's distribution with the cumulative density function

$$Q(x) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1} e^{-2j^2x^2}.$$

By Slutsky's theorem, $S \to 0$ when D_0 and D are congruent. Thus, given $b > 0$,

$$\delta = \frac{1}{1 + \exp\{a + b \times \log(S)\}} \to 1.$$

When D_0 and D are not congruent, since $S_{KS} = \max_{-\infty<t<\infty}\{|F(t) - G(t)|\}$, S_{KS} is bounded from 0. Thus, according to equation (4.10), $S \to \infty$ as $m \to \infty$, and thus $\delta \to 0$. ■

2. Calibration procedure for binary endpoints

We consider the case in which $D_0 = (x_1, \cdots, x_m)$ and $D = (y_1, \cdots, y_n)$ are distributed with $x_i \sim Ber(\mu_0)$ and $y_j \sim Ber(\gamma\mu_0)$, $i = 1, \cdots, m$ and $j = 1, \cdots, n$, where γ is the ratio or odds of the response rate between D and D_0. Given historical data D_0, the values of the tuning parameters a and b are calibrated as follows,

1. Estimate the response rate of D_0 by $\hat{\mu}_0 = \frac{\sum_{i=1}^{m} I(x_i)}{m}$ with $I(x_i)$ as the indicator function for counting the response.

2. Elicit from subject experts the maximum value of γ, denoted as γ_c, such that the difference between D and D_0 is practically negligible and they can be regarded as congruent, and the minimal value of γ, denoted as $\gamma_{\bar{c}}$, is such that D is deemed to be substantially different (i.e., not congruent) from D_0.

3. Generate R replicates of D by simulating $(y_1, \cdots, y_n)$ from $Ber(\gamma_c\hat{\mu}_0)$, and calculate the KS statistics between each of these R simulated data sets and D_0. Let $S^*(\gamma_c)$ denote the median of the R resulting KS statistics.

4. Repeat step 3 by replacing γ_c with $\gamma_{\bar{c}}$, and let $S^*(\gamma_{\bar{c}})$ denote the median of the R resulting KS statistics.

5. Solve a and b in (4) based on the two equations (4.5) and (4.6).

3. Evaluation of BBI

To implement the proposed design, we need to evaluate the BBI at each interim analysis, which depends on the posterior distributions of μ_T and μ_R. In what follows, we describe how to obtain these posterior distributions of μ_T and μ_R for evaluating the BBI at each interim. We first consider the case in which Y_T and Y_R are continuous endpoints following normal distributions $N(\mu_T, \sigma_T^2)$ and $N(\mu_T, \sigma_T^2)$, respectively. For

test arm T, we assign (μ_T, σ_T^2) Jeffrey's noninformative prior $f(\mu_T, \sigma_T^2) \propto \sigma_T^{-2}$, then given the interim data $D_T(n_k)$, the posterior distribution of μ_T is

$$f(\mu_T|D_T(n_k)) \sim t(\bar{y}_T,\ \hat{\sigma}_T^2/n_k,\ n_k - 1),$$

where $t(a, b, c)$ denote a t distribution with location parameter a, scale parameter b and degree of freedom c, and $\bar{y}_T$ and $\hat{\sigma}_T^2$ are the sample mean and variance of $D_T(n_K)$.

For the reference arm R, we employ the CPP approach to take advantage of the availability of historical data D_0. We assume the noninformative prior $f(\mu_R, \sigma_R^2) \propto \sigma_R^{-2}$ before observing D_0, and elicit δ following the CPP procedures by solving (4.7) and (4.8). Given the value of δ and interim data $D_R(n_k)$, the posterior of μ_R is given by

$$f(\mu_R|\delta, D_R(n_k)) \sim t\left(\frac{\delta m\bar{x_0} + n\bar{y}}{\delta m + n_k},\ \sqrt{\frac{2}{c(\delta)}\frac{1}{(\delta m + n_k - 2)(\delta m + n_k)}},\ \delta m - 2\right),$$

where $c(\delta) = 2/\{\delta m n_k(\bar{x_0} - \bar{y})^2/(\delta m + n_k) + \delta m\hat{\sigma_0^2} + n_k\hat{\sigma_R^2}\}$, $\bar{x}_0$ and $\bar{y}$ are the sample means of D_0 and $D_R(n_k)$, and $\hat{\sigma}_0^2$ and $\hat{\sigma}_R^2$ are the sample variances of D_0 and $D_R(n_k)$.

We now turn to the case in which Y_T and Y_R are binary and follow Bernoulli distributions $Ber(\mu_T)$ and $Ber(\mu_R)$, respectively. For arm T, we assign μ_T noninformative prior $Beta(1, 1)$, then the posterior of μ_T is given by

$$f(\mu_T|D_T(n_k)) \sim Beta(1 + \sum_{i=1}^{n_k} y_{T,i}, 1 + n_k - \sum_{i=1}^{n_k} y_{T,i}).$$

For arm R, starting from the noninformative prior $\mu_R \sim Beta(1, 1)$, we first apply the CPP approach to determine the value of δ. Given δ and $D_R(n_k)$, the posterior of μ_R is given by

$$f(\mu_R|\delta, D_R(n_k)) \sim Beta(\delta\sum_{i=1}^{m} x_i + \sum_{i=1}^{n_k} y_{R,i} + 1, \delta(m - \sum_{i=1}^{m} x_i) + (n_k - \sum_{i=1}^{n_k} y_{R,i}) + 1).$$

Here, y_T and y_R are realizations of Y_T and Y_R.

Table 4.1: The elicited values of a and b for CPP under 5 scenarios for normal endpoints

	Scenarios				
	1	2	3	4	5
$\hat{a}$	15.78	13.83	15.92	15.92	15.92
$\hat{b}$	6.18	5.59	6.22	6.22	6.22

Table 4.2: Simulation results of power and average sample size (n) for the normal endpoint with $\mu_R = 0$

	Historical data				μ_T				
Scenario	μ_0	N_0	Design		-0.223[a]	-0.115[b]	0[c]	0.115[d]	0.223[e]
			NB	Power	0.054	0.582	0.955	0.589	0.052
				n	76.98	93.32	85.31	93.26	76.41
1	0	300	CPP	Power	0.053	0.675	0.969	0.671	0.055
				n	76.21	88.79	76.51	88.54	76.13
			PP	Power	0.050	0.770	0.991	0.774	0.057
				n	75.76	83.36	60.20	83.88	75.80
		500	CPP	Power	0.052	0.690	0.976	0.692	0.054
				n	76.06	88.06	74.68	88.49	75.92
			PP	Power	0.041	0.697	0.980	0.694	0.054
				n	75.28	81.84	57.76	80.84	75.0
2	-0.5	300	CPP	Power	0.054	0.591	0.956	0.591	0.054
				n	76.94	94.04	85.24	93.00	76.21
			PP	Power	0.174	0.764	0.852	0.331	0.016
				n	87.8	89.96	83.24	83.48	60.24
		500	CPP	Power	0.053	0.587	0.958	0.597	0.052
				n	76.42	92.66	85.47	92.72	76.64
			PP	Power	0.220	0.728	0.825	0.313	0.009
				n	88.6	88.56	82.2	80.08	58.80
3	-0.3	300	CPP	Power	0.053	0.589	0.955	0.579	0.053
				n	76.70	93.37	85.22	93.23	76.33
			PP	Power	0.118	0.712	0.913	0.383	0.013
				n	87.44	95.04	86.8	88.12	65.4
		500	CPP	Power	0.054	0.582	0.955	0.589	0.050
				n	76.34	92.84	85.12	92.61	76.78
			PP	Power	0.141	0.751	0.926	0.366	0.022
				n	89.0	92.4	86.72	86.76	64.28
4	0.3	300	CPP	Power	0.053	0.587	0.936	0.591	0.056
				n	76.55	92.60	83.14	93.24	76.14
			PP	Power	0.004	0.135	0.689	0.893	0.277
				n	47.4	61.56	76.8	77.08	73.36
		500	CPP	Power	0.051	0.580	0.949	0.601	0.055
				n	76.49	92.64	85.32	93.69	76.37
			PP	Power	0.004	0.099	0.563	0.825	0.227
				n	46.28	56.64	70.72	78.32	71.92
5	0.5	300	CPP	Power	0.055	0.597	0.954	0.589	0.056
				n	76.71	93.56	85.01	93.37	76.86
			PP	Power	0.013	0.317	0.843	0.768	0.218
				n	59.48	79.84	82.12	89.56	88.40
		500	CPP	Power	0.052	0.589	0.953	0.584	0.055
				n	76.63	92.89	84.78	93.30	76.55
			PP	Power	0.012	0.297	0.823	0.771	0.187
				n	58.48	78.76	82.04	88.08	89.40

[a]Type I error rate
[b]Power
[c]Power
[d]Power
[e]Type I error rate

Table 4.3: Simulation results of power and average sample size (n) for the binary endpoint with $\mu_R = 0.5$

	Historical data				μ_T				
Scenario	μ_0	N_0	Design		0.4[a]	0.45[b]	0.5[c]	0.565[d]	0.625[e]
			NB	Power	0.045	0.615	0.939	0.589	0.05
				n	355.68	445.23	382.65	437.31	356.52
1	0.5	600	CPP	Power	0.05	0.711	0.966	0.718	0.05
				n	359.58	430.68	352.92	425.52	360.60
			PP	Power	0.045	0.674	0.982	0.703	0.03
				n	354	438.6	348.9	450.9	350.4
		1000	CPP	Power	0.047	0.73	0.972	0.767	0.048
				n	358.23	435.21	338.73	420.03	358.77
			PP	Power	0.042	0.713	0.978	0.686	0.045
				n	357.0	442.2	345.3	451.5	351.9
2	0.2	600	CPP	Power	0.05	0.612	0.935	0.588	0.046
				n	357.48	442.32	385.11	434.82	353.88
			PP	Power	0.152	0.66	0.820	0.416	0.018
				n	376.2	485.1	421.5	463.5	340.2
		1000	CPP	Power	0.048	0.616	0.935	0.585	0.049
				n	356.61	443.70	386.97	436.02	356.04
			PP	Power	0.175	0.656	0.926	0.454	0.015
				n	375.3	489.9	432.0	472.5	342.3
3	0.8	600	CPP	Power	0.043	0.612	0.940	0.593	0.047
				n	358.17	441.81	383.50	438.27	352.95
			PP	Power	0.027	0.524	0.943	0.636	0.144
				n	345.6	478.2	418.5	478.8	378.3
		1000	CPP	Power	0.048	0.623	0.941	0.593	0.046
				n	355.89	444.60	383.88	439.83	356.37
			PP	Power	0.019	0.525	0.933	0.629	0.174
				n	344.1	462.3	421.8	476.7	381.3
4	0.1	600	CPP	Power	0.044	0.625	0.939	0.593	0.046
				n	357.87	446.43	385.89	438.0	354.27
			PP	Power	0.152	0.655	0.921	0.493	0.019
				n	369.9	474	420	467.4	342.3
		1000	CPP	Power	0.043	0.633	0.941	0.598	0.048
				n	355.65	446.73	386.04	438.66	355.38
			PP	Power	0.242	0.634	0.925	0.490	0.026
				n	371.7	449.4	401.4	448.5	352.5
5	0.9	600	CPP	Power	0.049	0.620	0.938	0.596	0.05
				n	357.72	439.20	384.36	437.04	354.56
			PP	Power	0.029	0.520	0.947	0.575	0.147
				n	354.9	474.3	416.7	470.7	370.8
		1000	CPP	Power	0.047	0.621	0.942	0.595	0.048
				n	355.95	438.06	386.10	435.09	354.93
			PP	Power	0.029	0.519	0.933	0.584	0.148
				n	351.3	471.6	416.4	467.7	369.6

[a]Type I error rate
[b]Power
[c]Power
[d]Power
[e]Type I error rate

Table 4.4: Sensitivity analysis for the normal endpoint with $\mu_R = 0$ and $\sigma_R^2 = 0.25$

Historical data					μ_T				
μ_0	σ_0^2	N_0	Design		-0.223[a]	-0.115[b]	0[b]	0.115[b]	0.223[a]
0	1	500	CPP	Power	0.053	0.589	0.953	0.587	0.052
				n	76.94	93.98	86.45	94.08	76.8
			PP	Power	0.036	0.454	0.952	0.580	0.15
				n	80.0	98.4	87.6	98.2	79.3
0.3	4	500	CPP	Power	0.052	0.579	0.957	0.590	0.055
				n	76.55	93.62	85.69	92.12	76.44
			PP	Power	0.029	0.373	0.909	0.674	0.15
				n	74.6	95.84	91.68	100.28	84.92
0.5	1	500	CPP	Power	0.053	0.585	0.957	0.582	0.052
				n	77.02	93.47	85.37	92.95	76.93
			PP	Power	0.012	0.333	0.809	0.621	0.154
				n	60.8	89.4	91.3	100.16	88.96
0.5	4	500	CPP	Power	0.052	0.588	0.956	0.595	0.052
				n	76.15	93.88	85.55	93.23	76.63
			PP	Power	0.016	0.401	0.902	0.562	0.163
				n	71.68	92.08	90.2	101.08	87.36

a: Type I error rate; *b*: Power

Table 4.5: Application of the proposed CPP design to the biosimilar trial of Humira

Scenario	Clinical Response		Power		Sample Size	
	Humira	Test agent	CPP	NB	CPP	NB
1	0.608	0.608	81%	67%	190	188
2	0.486	0.486	76.4%	74.4%	256	259
3	0.608	0.486	4.3%	4.6%	140	137

Figure 4.1.: Power curve of the proposed CPP design for the normal endpoint when $\mu_0 = 0, 0.3$ and 0.5 and $\mu_R = 0$. The power curve of the NB design is shown as the reference.

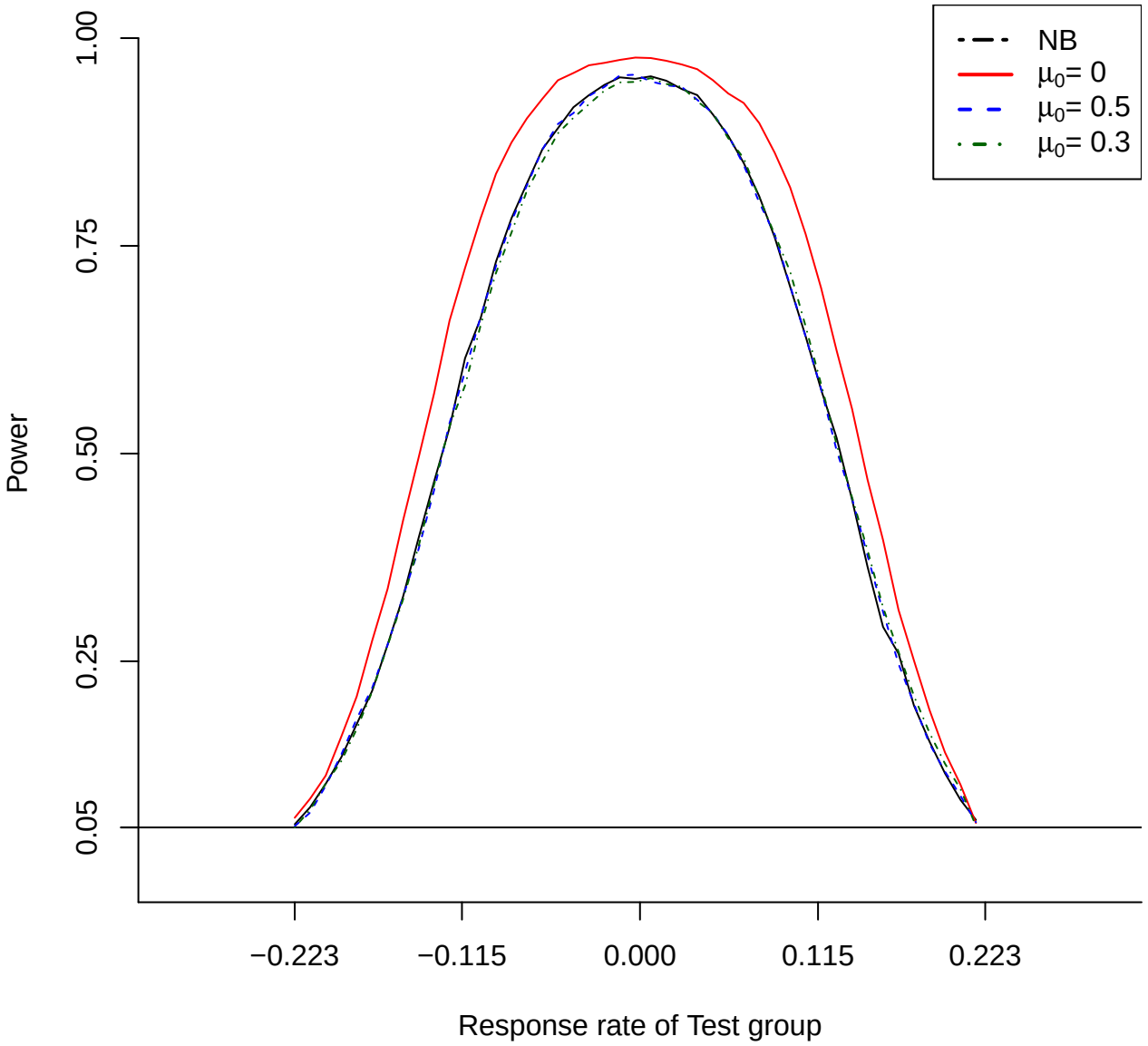
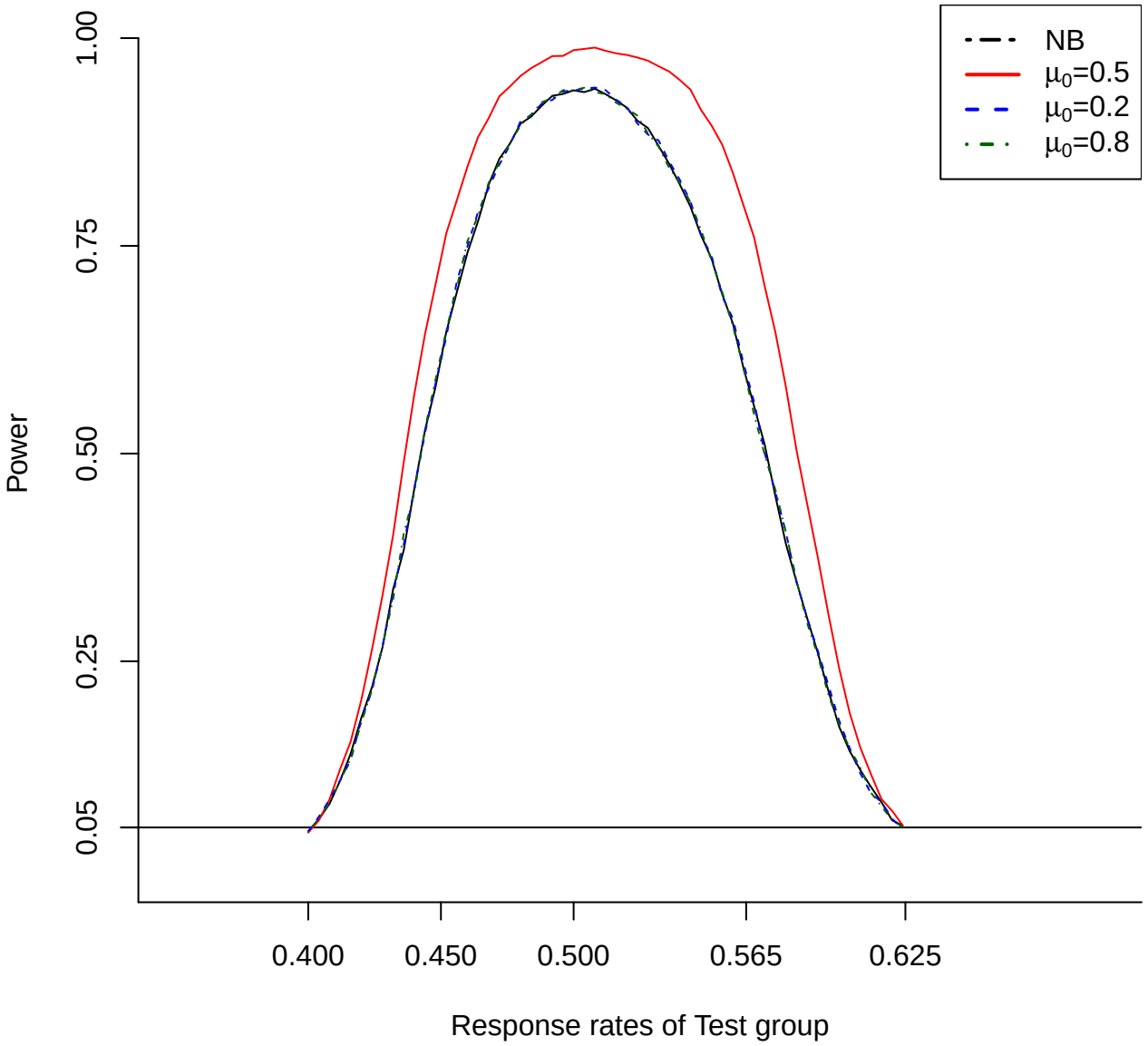


Figure 4.2.: Power curve of the proposed CPP design for the binary endpoint when $\mu_0 = 0.5, 0.2$ and 0.8 and $\mu_R = 0.5$. The power curve of the NB design is shown as the reference.



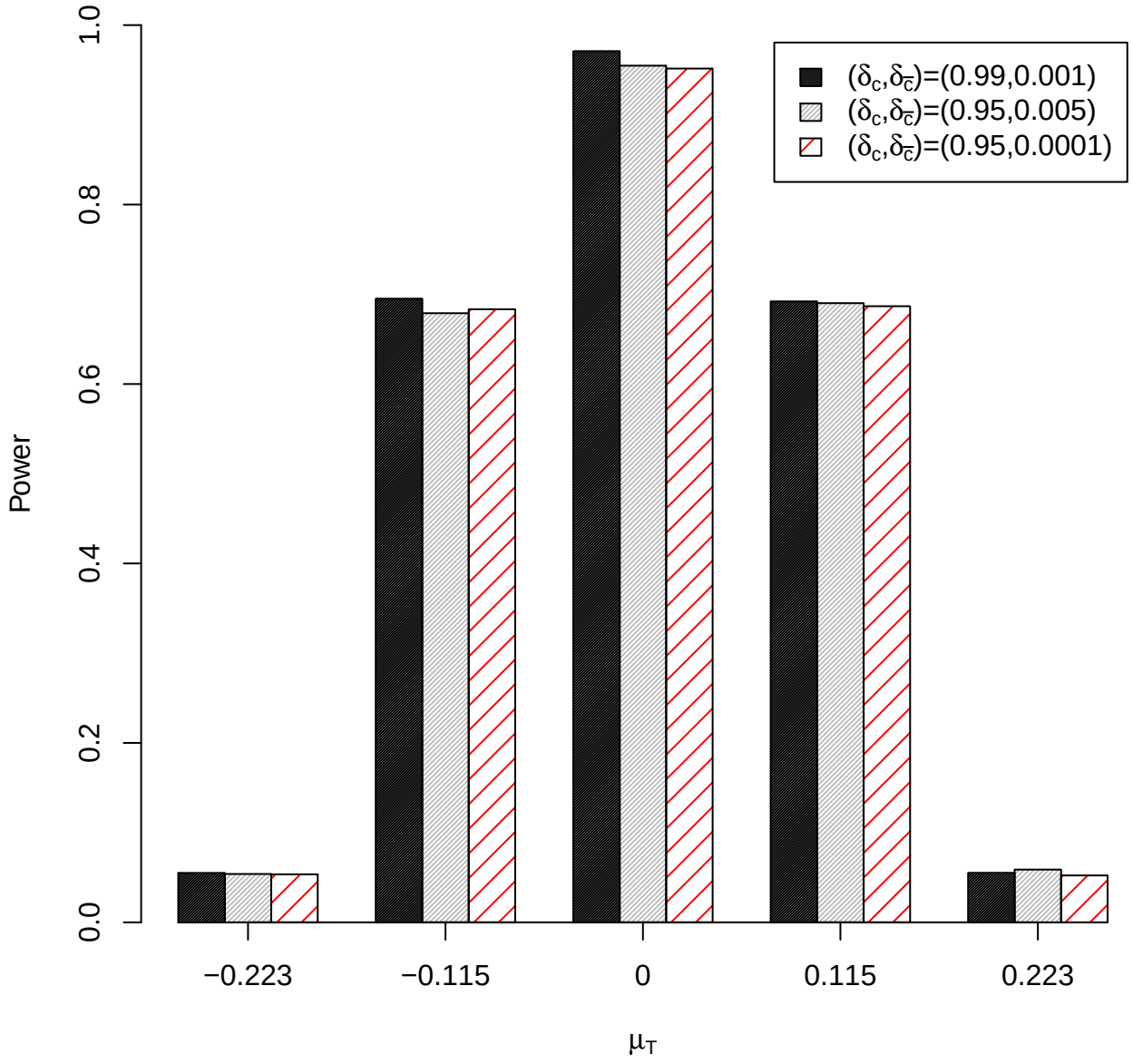


Figure 4.3.: Sensitivity analysis with different values of $(\delta_c, \delta_{\bar{c}})$ under scenario 1 with $N_0 = 500$.

5. Conclusions and Future Work

5.1 Conclusions

Several researches have shown that one main contribution to the high attrition of late phase is due to inefficiently conducting of the early phases, which is a main source for recommendation of under-therapeutic dose(s) to the subsequent confirmatory phase(s). This thesis mainly focuses on developing novel designs for phase I and II clinical trials.

In Chapter 2, we focus on adaptive phase I designs. Based on the BMA-CRM design (an extension of the CRM design), an automatic procedure has been proposed for prespecifying the multiple skeletons. By extensive simulation studies, the proposed method lets the BMA-CRM provide robust and better performance than the original version and has much higher performance than the CRM design. The algorithm has been developed into a Shiny app, which can be obtained freely online at *www.trialdesigns.org*.

In Chapter 3, we focus on an important design type, bridging trial, which gains special attention from drug administrative across nations and international pharmaceutical companies. The challenges of designing a bridge trial have at least two points, first is how to borrow the landmark trials' historical information of a certain ethnic populations into current study for a new ethnic population, second is how to acknowledge the heterogeneities among different ethnic populations. The CRM design framework has been used and that the toxic probabilities skeleton is a natural way to borrow the historical landmark trial information. A proposed mixture toxicity probability estimator is to estimate a dose-toxicity response curve, and from the available

discrete data on each dose level of landmark trial, a skeleton can be generated which then be used as a specified skeleton for the current study. This is the solution of how to efficiently borrow landmark information. Once we have one skeleton, we can shift this skeleton up and down to generate multiple skeletons, which obviously accommodate the true scenario that the landmark trial and the current trial should be heterogeneous but also not be deviated from each other so much. The shifting can be based on practitioner's experiences and properties of drugs or characteristics of the ethnic population. The BMA-CRM design introduced in chapter 2 is natural to incorporate multiple skeletons. The whole proposed structure above forms a novel bridge trial design, B-CRM design. The simulation studies demonstrate its desirable properties and a real trial for designing a phase I trial of the BKM120 in adult patients with advanced solid tumors has been shown to how to design a bridge trial and how to generate a clinical protocol in practice.

The above two researches deal with phase I trials. In Chapter 4, we focus on development of biosimilar products, which basically is a phase II trial problem with a feature of also existence of historical expired/to-be-expired reference drugs. Therefore, a main task is still to how to borrow historical information from the existing product. But, different from the above bridge trial, in which we borrow a "dose-toxicity curve" from landmark trial data, in biosimilar setting, the aim to conduct hypothesis testing, so we want to borrow information to enhance our understanding of a 'point' estimation. A calibrated power prior has been proposed to borrow information when the reference and historical arms are 'similar' to each other, and vice versa. A Bayesian similarity index is also proposed to assess congruence between the reference and historical arms. Simulation studies have been done to show that the proposed Bayesian group sequential design has better performance than a design without borrowing history information and another design with borrowing

history information but using the original power prior. A real trial for developing Adalimumab(Humira) biosimilar product has been shown as an example.

5.2 Future Work

Many unsettled issues in early phases are still open questions. Especially in cancer drug development area,

In Chapter 2, although the automatic algorithm for specifying multiple skeletons has been proposed, the method is essentially a computer-intensive way, which may be a daunting task for non-statistician to grasp the idea. Another issue is that the proposed algorithm lacks a theoretical framework. The optimal multiple skeletons specification problem is intrinsically an optimization problem, in which target argument function is PCS (percentage of correctly selection for MTD), different multiple skeletons produce different PCS values and that a certain multiple-skeletons achieves maximum PCS value is chosen as the optimal multiple-skeletons. A future work is to accomplish the proposed algorithm theoretical framework.

For designing trials of an agent for an ethnic population given a trial for another ethnic population had been conducted, in Chapter 3, a novel design, bridging CRM (BCRM) has been proposed. However, the ethnic populations can be generalized to more broader context, for instance, adult and pediatric populations, male and female patient populations, etc. Phase I trials for any kinds of sub-populations if they share biological or drug-responsive similarities can be conceptually designed using our proposed bridging design. An important area is to use the bridging idea for pediatrics trials since commonly clinical trials for adults have finished then initiation for children's trials. The challenge will be that dose range maybe very different between adult and child patients. Better use of the proposed should be discussed. This will be a future direction to modify the proposed method for pediatric drug development. Other fields including phase II and III bridging trials. The similar

idea for phase I can also be explored for late phases. For instance, in phase II dose-range trial, we still can model the trial's data to derive a dose-efficacy curve, based on response data on various doses, we can regard this curve as a 'skeleton-generated' curve, then the similar strategies of using BMA-CRM framework can be used naturally. However, for phase II and III hypotheses setting, the current method cannot be adopted straightforward, which will be future directions.

In Chapter 4, Bayesian group sequential design for biosimilar product with continuous and binary endpoints have been proposed. Therefore, design for time-to-event endpoint is a future work. Another interesting future work include: biosimilar phase I/II and II/III seamless designs under all kinds of endpoints scenarios.

Finally, all the topics covered in this thesis deal with single agent, especially for the previous two projects. Therefore, designs for combination trial for dual agents can use the proposed methods, but detailed discussion or extra modifications should be made in the future researches.

Bibliography

[1] Storer, B. E. (1989), "Design and Analysis of Phase I Clinical Trials," *Biometrics*, **45**, 925-937.

[2] O'Quigley, J., Pepe, M., and Fisher, L. (1990) Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, **46**, 33-48.

[3] Whitehead, J., and Brunier, H. (1995), "Bayesian Decision Procedures for Dose Determining Experiments," *Statistics in Medicine*, **14**, 885-893.

[4] Babb, J., Rogatko, A., and Zacks, S. (1998), "Cancer Phase I Clinical Trials: Efficient Dose Escalation With Overdose Control," *Statistics in Medicine*, **17**, 1103-1120.

[5] Leung, D., Wang, Y.G. (2001). An improved up-and-down design for Phase I trials . *Controlled Clinical Trials*, **22**, 126-138.

[6] Stylianou, M., and Flournoy, N. (2002), "Dose Finding Using the Biased Coin Up-and-down Design and Isotonic Regression," *Biometrics*, **58**, 171-177.

[7] Cheung, Y (2007) Sequential implementation of stepwise procedures for identifying the maximum tolerated dose *Journal of the American Statistical Association*, **102**:1448-1461.

[8] Ji Y, Li Y and Bekele B.N. (2007) Dose-finding in phase I clinical trials based on toxicity probability interval, *Clinical Trials*, **4**:235-244.

[9] Liu S and Yuan Y (2015) Bayesian Optimal Interval Designs for Phase I Clinical Trials, *Journal of the Royal Statistical Society: Series C*, **64(3)**: 507-523.

[10] Shen, L., and O'Quigley, J. (1996) Consistency of Continual Reassessment Method Under Model Misspecification, *Biometrika*, **83**, 395-405.

[11] Jia X, Cheung YK. (2014) Characterization of the likelihood continual reassessment method, *Biometriak*, **101**, 599-612.

[12] Yin G, Yuan Y. (2009) Bayesian model averaging continual reassessment method in phase I clinical trials. *Journal of the American Statistical Association*, **104**, 954-968.

[13] Daimon, T., Zohar, S., O'Quigley, J. (2011) Posterior maximization and averaging for Bayesian working model choice in the continual reassessment method. *Statistics in Medicine* 30, 1563-1573

[14] Asakawa, T., Hirakawa, A., Hamada, C. (2014) Bayesian model averaging continual reassessment method for bivariate binary efficacy and toxicity outcomes in phase I oncology trials *Journal of Biopharmaceutical Statistics*, 24, 310-325.

[15] Lee S.M and Cheung Y.K. (2009) Model calibration in the continual reassessment method. *Clinical Trials*, 6(3): 227-238.

[16] Cheung Y.K. (2011). *Dose Finding by the Continual Reassessment method*, Chapman & Hall, Boca Raton.

[17] Weisberg S. (2005). *Applied Linear Regression* 3rd Ed, John Wiley & Sons, Inc., Hoboken, New Jersey.

[18] Durham, S. D., Flournoy, N., and Rosenberger, W. F. (1997), A random walk rule for phase I clinical trials," *Biometrics*, **53**, 745-760.

[19] Chevret, S. (2006) *Statistical Methods for Dose-Finding Experiments*, John Wiley and Sons Ltd, England.

[20] Ting, N. (2006), *Dose Finding in Drug Development*, Springer, Cambridge, MA.

[21] Huang, S. and Temple, R. (2008) Is this the drug or dose for you? Impact and consideration of ethnic factors in global drug development, regulatory review, and clinical practice, *Nature*, **84**, 287-294.

[22] Yasuda, SU, Zhang, L and Huang S. (2008), The role of ethnicity in variability in response to drugs: focus on clinical pharmacology studies. *Nature*, **84**, 417-423.

[23] Barrera, K., Imagawa, D. K., Yamamoto, M., Katz, M., Sanati, H., Nguyen, A., Nguyen, L., Ho, J., Gebhardt, K. and Hoang, T. (2009) Sorafenib for treatment of hepatocellular carcinoma in the Asian-American population: Does one size fit all. 2009 ASCO Gastrointestinal Cancers Symposium (Abstract #198).

[24] Llovet, J., Ricci, S., Mazzaferro, V. , et al. (2008) Sorafenib in advanced hepatocellular carcinoma. *The New England Journal of Medicine*, **359**, 378-390.

[25] Joensuu H, Kellokumpu-Lehtinen PL, Bono P, Alanko T, Kataja V, Asola R, et al. (2006) Adjuvant docetaxel or vinorelbine with or without trastuzumab for breast cancer. *New England Journal of Medicine*, **354**, 809-820.

[26] Perng, R.P., Shih, J.F., Chen, Y.M., Chou, K.C., Lee, Y.C., Tsai, C.M. (2000) A phase II study of single agent docetaxel chemotherapy for non-small cell lung cancer. *Japanese Journal of Clinical Oncology*, **30**, 429-434.

[27] Muruyama R, Nishiwaki Y, Tamura T, Yamanoto N, Tsuboi M, Nakagawa K, et al. (2008) Phase III study, V-15-32, of gefitinib versus docetaxel in previously treated Japanese patients with non-small-cell lung cancer. *Journal of Clinical Oncology*, **26**, 4244-4252.

[28] Bruno R, Hille D, Riva A, Huinnink W, Van Oosterom A, Kaye S, et al. (1998) Population pharmacokinetics/ pharmacodynamics of docetaxel in phase II studies in patients with cancer. *Journal of Clinical Oncology*, **16**, 187-196.

[29] Goh BC, Lee SC, Wang LZ, F Lu, Guo JY, Lamba J, et al. (2002) Explaining interindividual variability of docetaxel pharmacokinetics and pharmacodynamics in Asians through phenotyping and genotyping strategies. *Journal of Clinical Oncology*, **20**, 3683-3690.

[30] Shih, W. J. (2001). Clinical trials for drug registration in Asian-Pacific countries: Proposal for a new paradigm from a statistical perspective. *Controlled Clinical Trials*, **22**, 357-366.

[31] Lan G. K. K., Soo Y. W., Siu C. T., Wang, M. (2005). The use of weighted Z-tests in medical research. *Journal of Biopharmaceutical Statistics*, **15**, 625-639.

[32] Gould, A. L., Jin, T., Zhang, L., Wang, W. (2012) A Predictive Bayesian Approach to the Design and Analysis of Bridging Studies. *Journal of Biopharmaceutical Statistics*, **22**, 916-934.

[33] Gandhi, M., Mukherjee, B., Biswas, D. (2012) A Bayesian Approach for Inference from a Bridging Study with Binary Outcomes. *Journal of Biopharmaceutical Statistics*, **22**, 935-951.

[34] Chow, S, Chieh, C., Liu, J. and Hsiao, C. (2012) Statistical Methods for Bridging Studies. *Journal of Biopharmaceutical Statistics*. **22**, 903-915.

[35] Morita, S. (2010) Application of the continual reassessment method to a phase I dose-finding trial in Japanese patients: East meets West. *Statistics in Medicine*. **30**, 2090-2097.

[36] Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. New York: Wiley.

[37] Ayer,M., Brunk, H.D., Ewing, G.M., Reid,W.T., Silverman, E. (1955) An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics* **26**, 641-647.

[38] Bhattacharya, R. and Kong, M. (2007) Consistency and asymptotic normality of the estimated effective doses in bioassay. *Journal of Statistical Planning and Inference*, **137**, 643-658

[39] Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression.* New York: Wiley.

[40] Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89** 1535-1546.

[41] Hoeting JA, Madigan D, Raftery AE and Volinsky CT (1999) Bayesian Model Averaging: A Tutorial *Statistical Science*, **14** 382-401.

[42] Garrett-Mayer, E. (2006) The continual reassessment method for dose-finding studies: a tutorial. *Clinical Trials*, **3**, 57-71.

[43] Lee, S. and Cheung, Y. (2009) Model calibration in the continual reassessment method. *Clinical Trials*, **6**, 227-238.

[44] Bendell, J.C., Rodon, J., Burris, H.A., Jonge, M., Verweij, J., Birle, D., Demanse, D., Buck, S.S., Ru, Q.C., Peters, M., Goldbrunner, M. and Baselga, J. (2012) Phase I, dose-escalation study of BKM120, an oral pan-class I PI3K inhibitor, in patients with advanced solid tumors, *Journal Of Clinical Oncology*, **30**, 282-290.

[45] Ishizuka N and Ohashi Y. (2001) The continual reassessment method and its applications: a Bayesian methodology for phase I cancer clinical trials. *Statistics in Medicine*, **20**, 2661-2681.

[46] Neuenschwander B., Branson M. and Gsponer T. (2008) Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in Medicine*, **27**, 2420-2439.

[47] FDA.(2012). Draft Guidance on Scientific Considerations in Demonstrating Biosimilarity to a Reference Product. U.S. Food and Drug Administration, Rockville, MD, U.S.A.

[48] FDA Basics. What is a biological product? http://www.fda.gov/AboutFDA/Transparency /Basics/ucm194516.htm.

[49] Datamonitor (2011). Pharmaceutical key trends 2011 - biosimilar market overview.

[50] Nemansky, M. (2014). Pharmacokinetic & Immunogenicity Bioanalysis in Biosimilar Development, *Biosimilars Newsletter* Vol 4.

[51] Ahn, C. and Lee, S-C. (2011). Statistical considerations in the design of biosimilar cancer clinical trials, *The Korean Journal of Applied Statistics.* 24(3): 495-503.

[52] Lin, J-R., Chow, S-C., Chang, C-H., Lin, Y-C., and Liu, J-P. (2013). Application of the parallel line assay to assessment of biosimilar products based on binary endpoints, *Stat. Med.* 32:449-461.

[53] Liao, J.Z., and Darken, P.F. (2013). Comparability of critical quality attributes for establishing biosimilarity, *Stat. Med.* 32:462-469.

[54] Zhang, N., Yang, J., Chow, S-C., Endrenyi, L., and Chi, E. (2013). Impact of variability on the choice of biosimilarity limits in assessing follow-on biologics, *Stat. Med.* 32:424-433.

[55] Yang, J., Zhang, N., Chow, S-C., and Chi, E. (2013). An adapted F -test for homogeneity of variability in follow-on biological products, *Stat. Med.* 32:415-423.

[56] Endrenyi, L, Chang, C., Chow, S-C., and Tothfalusi, L. (2013). On the interchangeability of biologic drug products, *Stat. Med.* 32:434-441.

[57] Chow, S-C., Endrenyi, L., and Lachenbruch, P.A. (2013). Comments on the FDA draft guidance on biosimilar products, *Stat. Med.* 32:364-369.

[58] Chow, S-C. (2013). Assessing biosimilarity and interchangeability of biosimilar products, *Stat. Med.* 32:361?363.

[59] Chow, S-C., Wang, J., Endrenyi, L., and Lachenbruch, P.A. (2013). Scientific considerations for assessing biosimilar products, *Stat. Med.* 32:370-381.

[60] Kang, S-H., and Chow, S-C. (2013). Statistical assessment of biosimilarity based on relative distance between follow-on biologics, *Stat. Med.* 32:382-392.

[61] Li, Y., Liu, Q., Wood, P., and Johri, A. (2013). Statistical considerations in biosimilar clinical efficacy trials with asymmetrical margins, *Stat. Med.* 32:393-405.

[62] Chow, S-C. (2013). *Biosimilars: Design and Analysis of Follow-on Biologics*, Chapman and Hall/CRC Biostatistics series; 60.

[63] Combest, A.J., Wang, S., Healey, B.T., and Reitsma, D.J. (2014). Alternative statistical strategies for biosimilar drug development, *Generic Biosimilars Init. J.* 3(1): 13-20.

[64] Kass, R.E. and Wasserman, L.A. (1996). The selection of prior distributions by formal rules, *J. Am Stat Assoc.* 91, 1343-1370.

[65] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall, London.

[66] Ibrahim, J.G., and Chen, M.H. (2000). Power prior distributions for regression models, *Stat. Sci.* 15,46-60.

[67] Duan, Y.Y., Smith, E.P., and Ye, K.Y. (2006). Using power priors to improve the binomial test of water quality, *J. Agric. Bio. Environ. Stat.* 11, 151-168.

[68] Neuenschwander B., Branson M., and Spiegelhalter, D.J. (2000). A note on the power prior, *Stat. Med.* 10;28(28):3562-6.

[69] Duan, Y.Y., and Smith, E.P. (2006). Evaluating water quality using power priors to incorporate historical information, *Environmetrics* 17:95-106.

[70] Duan, Y.Y., and Ye, K.Y. (2008). Normalized power prior Bayesian analysis, Working paper, The University of Texas at San Antonio.

[71] Strand, V. and Cronstein, B. (2014) Biosimilars: how similar? *Intern. Med. J.* 44:218-23.

[72] Chen, M.-H. and Ibrahim, J. G. (2000) Power prior distributions for regression models, *Stat. Sci.* 15:46-60.

[73] Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2003) On optimality properties of the power prior, *J. Am. Stat. Assoc.* 98:204-213.

[74] Chen, M.-H., Ibrahim, J. G., Shao, Q.-M., and Weiss, R. E. (2003) Prior elicitation for model selection and estimation in generalized linear mixed models, *J. Stat. Plan. Inference* 111:57-76.

[75] Smirnov, N. V. (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. (Russian) *Bull. Moscow Univ.* 2:3-16.

[76] Thall, P. and Simon, R. (1994) Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 50, 337-349.

[77] Yuan, Y. and Yin, G. (2009) Bayesian dose-finding by jointly modeling toxicity and efficacy as time-to-event outcomes. *J. R. Stat. Soc. Ser. C Appl. Stat.* 58, 719-736.

[78] Silman AJ, Hochberg MC. Epidemiology of the rheumatic diseases. 2nd edition: Oxford University Press; 2001.

[79] E.C. Keystone, A F. Kavanaugh, J.T. Sharp, Hyman Tannenbaum, Ye Hua, Leah S. Teoh, S. A. Fischkoff, and Elliot K. Chartash. (2004) Radiograhic, Clinical, and Functional outcomes of treatment with Adalimumab (a human anti-humor necrosis factor monoclonal antibody) in patients with active rheumatoid arthritis receiving concomitant methotrexate therapy. *Arthritis & Rheumatism* 50(5): 1400-1411.

[80] Haitao Pan & Ying Yuan. (2016) A default method to specify skeletons for Bayesian model averaging continual reassessment method for phase I clinical trials. *Statistics in Medicine* 36(2): 266-279.

[81] Suyu Liu, Haitao Pan, Jielai Xia, Qin Huang & Ying Yuan. (2015) Bridging continual reassessment method for phase I clinical trials in different ethnic populations. *Statistics in Medicine* 34(10): 1681-1694.

[82] Haitao Pan, Ying Yuan & Jielai Xia. (2017) A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials. *J. R. Stat. Soc. Ser. C Appl. Stat.* DOI: 10.1111/rssc.12204.

Vita

Haitao Pan was born in Xianyang, Shaanix, China on April, 1978, the son of Shilin Pan and Yongli Xi. After receiving the Ph.D degree with a major in Preventive Medicine from Fourth Military Medical University in May, 2012, I worked as a research lecturer in the Department of Statistics at Xi'an University of Finance and Economics. In August of 2014, I entered The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.

Permanent address:

Chongye Rd, Building 4, Yanta District

Xi'an, China, 710061