

5-2020

BAYESIAN ADAPTIVE DESIGNS AND VERSATILE SOFTWARE PLATFORM FOR EARLY PHASE CLINICAL TRIALS

Yanhong Zhou

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations



Part of the [Medicine and Health Sciences Commons](#)

Recommended Citation

Zhou, Yanhong, "BAYESIAN ADAPTIVE DESIGNS AND VERSATILE SOFTWARE PLATFORM FOR EARLY PHASE CLINICAL TRIALS" (2020). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 996.
https://digitalcommons.library.tmc.edu/utgsbs_dissertations/996

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

BAYESIAN ADAPTIVE DESIGNS AND VERSATILE SOFTWARE PLATFORM FOR EARLY PHASE CLINICAL TRIALS

by

Yanhong Zhou, M.S.

APPROVED:



J. Jack Lee, Ph.D.
Advisory Professor



Ying Yuan, Ph.D.
Co-advisory Professor



Xuelin Huang, Ph.D.



Ruitao Lin, Ph.D.



Luis Gonzalo Leon Novelo, Ph.D.



Chad Tang, M.D.

APPROVED:

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

BAYESIAN ADAPTIVE DESIGNS AND VERSATILE SOFTWARE PLATFORM FOR EARLY PHASE CLINICAL TRIALS

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

by

Yanhong Zhou, M.S.

Houston, Texas

May, 2020

To the Lord our God who made it all possible

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest appreciation to my advisors Dr. J. Jack Lee and Dr. Ying Yuan. They are to me what compass is to sailors. When I sail in the wrong way on the sea of research, they show me the right direction. With their guidance and supervision, I proceed with much more confidence and arrive at my target destinations much quicker. They are to me also what parents are to their kids. They always want the best for me: unconditionally supporting me for any adventure I want to take; tirelessly mentoring me in professional development (e.g., presentation, writing, and attending conferences); continuously trusting me even when I did not (e.g., when I would fail at a project); and greatly empowering me to be an independent and reliable researcher. Many thanks also extend to the rest of my advisory committee: Dr. Xuelin Huang, Dr. Luis Gonzalo Leon Novelo, Dr. Chad Tang, and Dr. Ruitao Lin. They have provided many insightful ideas and helpful feedbacks for my research. A special thank you goes to Dr. Ruitao Lin, who provided much advice even before joining the advisory committee.

Besides my advisory committee, my sincere thanks also go to our wonderful GSBS leadership and kindhearted administration team. Dr. Bill Mattox, Dr. Prahlad Ram, and Dr. Liang Li have provided their continuous help in all aspects of my education training at GSBS. Along with their exemplary leadership, our administration team (e.g., kindhearted Ms. Brenda Gaughan and responsible Ms. Amy Carter, among others), have dedicated their time to helping me navigate through

any milestone in the past four years. In addition, I would like to acknowledge Ms. Jessica Swan for her great help with editing the manuscripts I co-authored and Dr. Ying-Wei Kuo for providing great advice in software development.

I am also very grateful to my colleagues who had greatly helped me at the beginning of my Ph.D. study: Dr. Heng Zhou, Dr. Liangcai Zhang, Dr. Haitao Pan, and Dr. Yiyi Chu. For example, Drs. Zhou and Zhang have been always available to answer my “silly” questions and provided their help whenever they could. I also highly appreciate the friendship from Dr. Youyi Zhang, Dr. Tianzhong Yang, Dr. Fang Wang, and my fellow schoolmates: Wen Li, Na Liu, Xinyue Qi, Shikun Wang, Ziqiao Wang, Yi Yao. They have either spent their precious time to be with me talking about life or to discuss our course works throughout my PhD study.

More importantly, I want to thank my family for their support in the past four years. I want to thank my mother-in-law, Fen Zhang, for her great motherhood that profoundly benefits me. She has treated me like her own daughter and provided unconditional love to support so that I could pursue my education while I am taking care of my daughter. She came to help when I prepared for my candidacy exam and even planned to come to help while I was writing my dissertation (but she couldn’t because of the outbreak of the COVID-19). My husband Feng is also very supportive and collaborates with me to better balance our duties between family and work. My daughter Angela is an angel from our God. She is sweet and also very cooperative, especially when her school is closed during this particularly hard time.

Last but not least, I sincerely thank our God for strengthening me in my hard times, supplying endless joy and inexhaustible wisdom, and making the impossible possible.

ABSTRACT

BAYESIAN ADAPTIVE DESIGNS AND VERSATILE SOFTWARE PLATFORM FOR EARLY PHASE CLINICAL TRIALS

Yanhong Zhou, M.S.

Advisory Professors: J. Jack Lee, Ph.D. & Ying Yuan, Ph.D.

This research is dedicated to improve the efficiency of Bayesian adaptive designs for early phase clinical trials. Phase I clinical trials can be conducted using algorithm-based (also known as rule-based), model-based, or model-assisted designs. Numerous studies have shown that model-assisted designs have the simplicity of algorithm-based designs while possessing the great performance of the model-based designs. Despite the desirable properties of current model-assisted designs, their use is still limited. More importantly, they can fall short of tackling some emerging challenges brought by the development of novel therapies. Thus, there is a pressing need for model-assisted designs that can complement the current designs. In this work, we first clarify some mis-conceptions between algorithm-based designs and the model-assisted designs with the purpose to eliminate the confusion caused by the designs' similarity in their appearance. Second, we develop a class of novel model-assisted designs that aim to accommodate the urgent need to utilize readily available historical data or real-world evidence to further improve the efficiency of the Phase I model-assisted designs. Third, we construct a seamless Phase I/II design that addresses the challenges emerging along with the vast development of immunotherapy and targeted therapy. Fourth, we develop a versatile software platform to provide user-friendly web-based

applications to facilitate the use of a series of well-performed model-assisted designs that are built on sound statistical foundations and have superior operating characteristics (e.g., have high probability of identify the MTD and treat a large number of patients on the MTD). In addition to the important issues addressed for Phase I model-assisted designs, we also include the examination of a critical topic in Phase II clinical trials: sequential monitoring. We thoroughly study the connections between different sequential monitoring approaches theoretically. Furthermore, we conduct extensive simulations to examine the impact of different types of prior distributions on the false positive rate and power to test the efficacy of a treatment, and provide practical recommendations for Phase II sequential monitoring. Our research will greatly advance drug development as it not only provides a wide range of innovative designs, but also creates user-friendly versatile software platforms to facilitate the implementation of the novel designs.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	vi
List of Figures	xiv
List of Tables	xx
 CHAPTER	
1. Background	1
1.1 Model-assisted design versus algorithm-based design	5
1.2 Use informative prior to improve the efficiency of model-assisted Phase I clinical trial designs	6
1.3 Seamless Phase I/II design for identifying optimal biological dose	7
1.4 Versatile software platform of Bayesian adaptive designs for early phase clinical trials	9

1.5	Sequential monitoring in Phase II designs	11
2.	A comparative study of Bayesian optimal interval (BOIN)	
	design with interval 3+3 (i3+3) design for Phase I oncology	
	dose-finding trials	15
2.1	Introduction	15
2.2	Method	17
2.2.1	Bayesian optimal interval design	17
2.2.2	Interval 3+3 design	21
2.2.3	Comparison of BOIN and i3+3 designs	22
2.3	Simulation study	24
2.3.1	Simulation configuration	24
2.3.2	Results	25
2.3.3	The role of “de-escalation modification rule”	28
2.4	Summary	28
A	Appendix	31
A2.0.1	A subset of random scenarios when comparing BOIN	
	to i3+3	31
A2.0.2	The simulation results for BOIN and i3+3 designs	
	under 1000 random scenarios	32
3.	Incorporating historical information to improve phase I clin-	
	ical trial designs	34

3.1	Introduction	34
3.2	Method	36
3.2.1	Incorporate prior information in CRM	36
3.2.2	Incorporate prior information in BOIN	38
3.2.3	Incorporate prior information in keyboard/mTPI-2 design	42
3.2.4	Robust prior	44
3.2.5	Choose PESS	45
3.3	Software	46
3.4	Simulation	46
3.4.1	Fixed scenarios	46
3.4.2	Random scenarios	51
3.4.3	Unequal prior information across doses	54
3.5	Summary	55
A	Appendix	57
A3.0.1	Determining informative prior for BOIN	57
A3.0.2	iBOIN Shiny app interface	58
A3.0.3	Random scenario configuration	59
A3.0.4	Generate random scenarios where prior MTD is cor- rectly specified	59
A3.0.5	Generate random scenarios with different levels of mis-specification	61

4. A utility-based Bayesian optimal interval (U-BOIN) Phase I/II design to identify the optimal biological dose for targeted and immune therapies	62
4.1 Introduction	62
4.2 Method	65
4.2.1 Efficacy-toxicity model	65
4.2.2 Utility	66
4.2.3 Optimal biological dose (OBD)	68
4.2.4 Phase I/II OBD finding algorithm	69
4.2.5 Delayed efficacy response	72
4.2.6 Software and trial implementation	76
4.3 Numerical study	78
4.3.1 Simulation A	78
4.3.2 Simulation B	82
4.3.3 Sensitivity analysis	84
4.4 Summary	85
A Appendix	89
 5. Versatile software platforms for the implementation of novel model-assisted designs in early phase clinical trials	 90
5.1 Introduction	90
5.2 Method	93
5.2.1 BOIN suite	93

5.2.2	Keyboard suite	100
5.2.3	Tools for software development	103
5.3	Results	104
5.3.1	Features of the applications	104
5.3.2	Implementation with a trial example	106
5.4	Summary	109
A	Appendix	110

**6. The use of local and nonlocal priors in Bayesian test-Based
monitoring for Phase II clinical trials 117**

6.1	Introduction	117
6.2	Sequential monitoring	120
6.2.1	Bayesian posterior (or predictive) probability based (PB) monitoring	120
6.2.2	Bayesian hypothesis test based (TB) monitoring . .	121
6.2.3	Relationship between PB monitoring and TB mon- itoring	123
6.3	Prior specification	124
6.3.1	Beta distribution	124
6.3.2	The inverse moment (iMOM) density	125
6.3.3	Prior effective sample size	125
6.4	Simulation	127
6.4.1	Simulation setting	127
6.4.2	Simulation results	128

6.5	Software application	131
6.6	Summary	134
A	Appendix	135
	A6.0.1 Interpretation of Bayes factor as the strength of ev- idence	135
	A6.0.2 iMOM density with varying parameters	135
	A6.0.3 The relative robustness of iMOM versus Beta	137
	A6.0.4 Sensitivity analysis	137
	A6.0.5 Output for the shiny application example	144
7.	Conclusion	146
	Bibliography	150
VITA	161

List of Figures

2.1	Dose escalation and de-escalation algorithm in BOIN and i3+3. DLT rate is the observed DLT probability. The main difference between two designs is that i3+3 adds a “de-escalation modification rule” (in red).	18
2.2	Performance difference between BOIN and i3+3 under 1000 randomly generated scenarios. A positive value for “Correct selection (%)” and “Number of patients treated on MTD” indicates that BOIN outperforms i3+3, while a negative value of “Patients treated above MTD (%)” indicates that BOIN is safer.	25
2.3	Performance difference between i3+3s (i.e., the counterpart of i3+3 with the “de-escalation modification rule” removed) and i3+3 under 1000 random scenarios. A positive value for “Correct selection (%)” and “Number of patients treated on MTD” indicates that i3+3s outperforms i3+3, while a negative value of “Patients treated above MTD (%)” indicates that i3+3s is safer.	29
A2.1	A sample of 50 randomly generated scenarios given the target DLT probability $\phi=0.2, 0.3$ or 0.4	31

A2.2	Performance of BOIN and i3+3 designs under 1000 randomly generated scenarios given different target DLT probability (ϕ) and cohort size.	33
3.1	Escalation and de-escalation boundaries (λ_e, λ_d) of iBOIN given different prior DLT probability (q) and PESS = 3 or 5, in comparison to the boundaries determined using non-informative prior in standard BOIN.	41
3.2	Operating characteristics of iCRM, iBOIN, and iKeyboard, in comparison to their counterparts with non-informative priors, under 2000 random scenarios when the prior is <i>correctly specified</i> . iBOIN _R and iKeyboard _R are iBOIN and iKeyboard using robust prior, respectively. The number under each boxplot is the average value with the standard deviation shown in parenthesis.	52
3.3	Operating characteristics of iCRM, iBOIN and iKeyboard, in comparison to their counterparts with non-informative priors, under 4000 random scenarios when the prior MTD is <i>one dose off</i> the true MTD. iBOIN _R and iKeyboard _R are iBOIN and iKeyboard using robust prior, respectively. The number under each boxplot is the average value with the standard deviation shown in parenthesis.	53
3.4	Operating characteristics of iCRM, iBOIN and iKeyboard, in comparison to their counterparts with non-informative priors, under 4000 random scenarios when the prior MTD is <i>two doses off</i> the true MTD. iBOIN _R and iKeyboard _R are iBOIN and iKeyboard using robust prior, respectively. The number under each boxplot is the average value with the standard deviation shown in parenthesis. . .	54

3.5	Operating characteristics of iBOIN and iCRM when different amount of prior information (i.e., PESS) is available for different doses under scenarios 1 to 5.	55
A3.1	User interface of iBOIN software	58
A3.2	50 randomly selected scenarios from the 2000 scenarios generated	60
4.1	Diagram of the U-BOIN design	70
4.2	The user interface of the U-BOIN software	77
4.3	Simulation scenarios. The dash-dotted line (blue) is the dose-toxicity curve, the solid line (red) is the dose-efficacy curve, and the dashed line (black) is the dose-immune response curve. The OBD is highlighted by red asterisk in the x-axis. Simulation A considers only the efficacy and toxicity curves, while simulation B considers efficacy, toxicity, and immune response.	79
4.4	Results of sensitivity analysis for different utilities. Scenario 8 is not included, as the OBD does not exist in that scenario.	86
4.5	Results of sensitivity analysis for different patient allocation strategies: pick-the-winner (PW), adaptive randomization (AR), and equal randomization (ER). Scenario 8 is not included, as the OBD does not exist in that scenario.	87
5.1	The waterfall design for finding the MTD contour in drug-combination trials (adapted from Figure 1 in Zhang and Yuan [121]). For panels (a)-(c), the circles surrounded by a rectangle is a subtrial. In panels (b)-(c), the filled circle is the MTD candidate selected from the previous subtrial. The MTD candidate for the third subtrial in panel (c) is indicated in the third row of panel (d).	99

5.2	Illustration of dose escalation and de-escalation in keyboard design .	101
5.3	Bayesian optimal interval (BOIN) and Keyboard suite available at www.trialdesign.org.	105
5.4	Design setting and flowchart for the trial example using the BOIN app.	107
5.5	Decision table for the trial example using the BOIN app.	108
5.6	Simulation example using the BOIN app.	108
5.7	Design setting of trial example and flowchart produced in the BOIN app.	116
6.1	Efficacy and futility sequential monitoring using Bayes factor with weakly informative prior (ESS=5, left panels) and strongly informa- tive prior (ESS=20, right panels). When ESS=5, the iMOM has parameters $k = 0.325$, $\nu = 0.65$, $\tau = 0.7029$ and the local prior is Beta(2.2, 2.8). When ESS=20, the iMOM has parameters $k = 1.685$, $\nu = 3.37$, $\tau = 0.0467$ and the local prior is Beta(8.2, 11.8). The max- imum sample size is 50.	130
6.2	Probabilities of claiming efficacy when H_1 is true (i.e., panel a), and probabilities of claiming futility when H_0 is true (i.e., panel b) at each interim time and final analysis when maximum sample size N = 50 and prior effective sample size is 5 . The numbers in parentheses represent the overall probabilities of claiming futility or efficacy in the trial.	131
6.3	Shiny app interface of the TB monitoring procedure based on the iMOM prior.	132

A6.1	iMOM density with a null value at 0.2 under different parameter (k, τ, ν) settings.	136
A6.2	Plot of prior, likelihood, and posterior distribution for the setting described in [25]. All the prior distributions have the same mode. The prior mean response rate is 0.2, corresponds to a prior log(odds) of -1.39 , but the observed response rate is 0.8 with a sample size of 20.	137
A6.3	Efficacy and futility sequential monitoring using Bayes factor with weakly informative prior distributions (ESS=5) and a strongly infor- mative prior (ESS=20). When ESS=5, the iMOM has parameters $k = 0.325$, $\nu = 0.65$, $\tau = 0.7029$ and the local prior is Beta(2.2, 2.8). When ESS=20, the iMOM has parameters $k = 1.685$, $\nu = 3.37$, $\tau = 0.0467$ and the local prior is Beta(8.2, 11.8). The maximum sample size is 100.	139
A6.4	Probability of claiming efficacy under H_1 and probability of claiming futility under H_0 at each interim or final analysis when maximum sample size $\mathbf{N} = \mathbf{50}$ and prior effective sample size is $\mathbf{20}$. The num- bers in parentheses represent the overall probabilities of claiming futility or efficacy in the trial.	140
A6.5	Probability of claiming efficacy under H_1 and probability of claiming futility under H_0 at each interim or final analysis when maximum sample size $\mathbf{N} = \mathbf{25}$ and prior effective sample size is $\mathbf{5}$. The numbers in parentheses represent the overall probabilities of claiming futility or efficacy in the trial.	141

A6.6	Probability of claiming efficacy under H_1 and probability of claiming futility under H_0 at each interim or final analysis when maximum sample size $\mathbf{N} = \mathbf{100}$ and prior effective sample size is $\mathbf{5}$. The num- bers in parentheses represent the overall probabilities of claiming futility or efficacy in the trial.	142
A6.7	Probability of claiming efficacy under H_1 and probability of claiming futility under H_0 at each interim or final analysis when maximum sample size $\mathbf{N} = \mathbf{100}$ and prior effective sample size is $\mathbf{20}$. The numbers in parentheses represent the overall probabilities of claiming futility or efficacy in the trial.	143
A6.8	Full stopping boundaries and iMOM prior determined by the input under Trial setting in the Shiny app, as shown in Figure 6.3. . . .	144
A6.9	Example of simulation output.	145
A6.10	An example for the template of protocol.	145

List of Tables

2.1	The optimal escalation/de-escalation boundaries (λ_e, λ_d) under the BOIN design for different target DLT probabilities.	18
2.2	Fifteen representative randomly generated scenarios	26
2.3	Operating characteristics of BOIN and i3+3 under the 15 representative scenarios shown in Table 2.2.	27
3.1	iBOIN decision boundaries up to 30 patients with a cohort size of 3, given the skeleton $(q_1, \dots, q_5) = (0.10, 0.19, 0.30, 0.42, 0.54)$ and PESS $n_{01} = \dots = n_{05} = 3$. The target DLT probability $\phi = 0.3$. . .	42
3.2	Ten dose-toxicity scenarios with target DLT probability $\phi = 0.30$. The prior MTDs are correctly specified in scenarios 1-5 and misspecified in scenarios 6-10.	47
3.3	Operating characteristics of iCRM, iBOIN and iKeyboard, in comparison with their counterparts with non-informative priors. iBOIN _R and iKeyboard _R are iBOIN and iKeyboard using robust prior. . . .	48
4.1	Examples of utility	67
4.2	Dose escalation and de-escalation boundaries of the BOIN design .	70

4.3	Results of Simulation A, including the selection percentage (Selection %), the average number of patients treated at each dose (No. of patients), and the percentage of early stopping. The optimal biological dose (OBD) is bolded. In scenario 8, the OBD does not exist, and thus the percentage of early stopping is bolded.	80
4.4	Results of Simulation B, including the selection percentage (selection %), the average number of patients treated at each dose (No. of patients), the percentage of early stopping, and the trial duration. The optimal biological dose (OBD) is bolded. In scenario 8, the OBD does not exist, and thus the percentage of early stopping is bolded.	83
5.1	Dose escalation/de-escalation and elimination boundaries in BOIN design with target DLT probability of 0.3	94
6.1	Full stopping boundaries for efficacy and futility monitoring using Bayes factor. The null response rate is 0.2 and the alternative response rate is 0.4. The trial stops for futility if $BF_{01} > 9$ and for efficacy if $BF_{10} > 9$. The prior effective sample size for the iMOM prior is 10. Patients enter the trial with a cohort size of five and the first interim is conducted after ten patients are treated.	133

6.2	Operating characteristics for a phase II study with sequential monitoring for both efficacy and futility using Bayes factor with iMOM prior. The null response rate is 0.2 and the alternative response rate is 0.4. The trial stops for futility if $BF_{01} > 9$ and for efficacy if $BF_{10} > 9$. The prior effective sample size for the iMOM prior is 10. Patients enter the trial with a cohort size of five. The first monitoring is conducted after ten patients are treated and subsequent interim monitoring is carried out after every five patients. The maximum number of patients enrolled is 50.	134
A6.1	Interpretation of Bayes factor(BF_{10}) and its log scale as the strength of evidence for H_1 over H_0 [47]	135

CHAPTER 1

Background

A clinical trial is defined by the National Institutes of Health (NIH) as a research study in which human subjects are prospectively assigned to receive an intervention (e.g., a new drug) in order to evaluate the effect of this intervention on “health-related biomedical or behavioral outcomes” [76]. In a clinical trial, the outcome used to objectively measure the effect of the intervention is referred to as endpoint. Common endpoints are dose limiting toxicity (DLT), tumor response, and survival.

Traditionally, clinical trials for drug development involve several critical stages: Phase I, Phase II, Phase III, and Phase IV, each with a different objective. A Phase I trial primarily aims at identifying the maximum tolerated dose (MTD), defined as the highest dose with acceptable level of toxicity. After the toxicity profile has been investigated and the MTD for the new treatment has been determined, a Phase II clinical trial is conducted at the MTD to evaluate whether the new treatment is efficacious. Once the efficacy of the drug is confirmed in a Phase II study based on pre-selected short-term endpoint (e.g., tumor response), a confirmatory trial (i.e., Phase III) will be conducted. In Phase III, the most desirable outcome is to get the drug approved by relevant regulatory agents (e.g., Food and Drug Administration in

the U.S.). The primary endpoint is often time to a certain adverse event (e.g., death) and the approval of the drug requires comparing the new treatment to a standard of care with a certain improvement margin (e.g., on overall survival). Once the drug is approved based on the results of a Phase III study and released into the market, a Phase IV trial may be conducted to continuously monitor the long-term side effect of the drug.

Our research work is primarily focused on Bayesian adaptive designs for early phase (Phase I or II) clinical trials. The main purpose of Phase I studies is to evaluate the safety (or study the toxicity profile) of a new drug and identify the MTD. In general, the endpoint in Phase I trials is binary, i.e., whether or not a patient experiences a dose-limiting toxicity (DLT), which refers to side effects caused by the treatment under investigation that are severe enough to prevent an increase in the dosage of the drug. The corresponding parameter of interest (i.e., any numerical quantity that tells us something about the effect of the new intervention) is the proportion of patients who experience at least one DLT. The earliest class of designs proposed to identify the MTD is rule-based [4] or algorithm-based [119]. Rule or algorithm-based means that there is no assumption on the true dose-toxicity curve and that dose assignment in the trial is based on a set of ad hoc prespecified rules. Classical examples include the conventional 3+3 design originated in the 1940s [19], the pharmacologically guided dose escalation design [17], and the accelerated titration design [89] proposed in the 1990s, among others. The eminent advantages of the algorithm-based designs are that they are easy to implement and require no special software. However, they are typically rigid and also have undesirable properties, such as the implicit target DLT probability, poor accuracy to identify the MTD, and the likeliness of assigning patients to sub-therapeutic doses [114].

Other good alternatives for finding the MTD are model-based designs. Model-based designs assume that dose-toxicity curve is monotonically increasing with dose levels. The most widely used or cited model-based design are continual reassessment method (CRM [77]) and its extensions: escalation with overdose control (EWOC [2]), Bayesian logistics regression model (BLRM [74]), and Bayesian model averaging CRM (BMA-CRM [111]), among others [53, 63, 68]. Compared to algorithm-based designs, the model-based designs generally have better operating characteristics, including greater accuracy to identify the MTD and a larger number of patients assigned to the MTD [119]. However, because the model-based designs require repeat model fitting during trials to guide dose assignment, they are perceived to be “black boxes” that are hard for practitioners to comprehend. Due to such a perception along with the inconvenience of the implementation, their use is quite limited. Thus, despite the conventional 3+3 design has well-documented flaws, it remains dominant in Phase I clinical trials.

To combine the simplicity of algorithm-based designs and the superior performance of model-based designs, a series of model-assisted designs have been proposed, including the Bayesian optimal interval design (BOIN [65]), keyboard [105], modified toxicity probability interval design (mTPI [43]), among a few others ([31, 60]). Similar to the model-based designs, the model-assisted designs utilize a statistical model to derive efficient decision rules for real-time decision making. The difference between model-assisted and model-based designs is that the dose escalation and de-escalation rules in the former can be pre-determined before the onset of a trial. Thus, model-assisted designs are as simple as the conventional 3+3 to implement. Many studies have conducted comprehensive comparisons among algorithm-based, model-based, and model-assisted designs. The studies found that the easy-to-use model-assisted

designs greatly outperform the 3+3 design and have comparable performance to the more complicated model-based designs [85, 124, 125]. Among the model-assisted designs, BOIN and keyboard stand out with their greater accuracy, safety, and reliability. Here, accuracy is quantified using the probability of correctly identifying the MTD and the number of patients treated at the MTD; safety is measured by the percentage of treating patients above MTD and of selecting overly toxic dose as MTD; and reliability is measured by the probability of assigning 50% or more patients at doses above the MTD and the probability of assigning less than six patients at the MTD.

Despite the marvelous performance of the model-assisted designs, the designs can sometimes fall short of tackling the emerging challenges in recent early phase trials. Additionally, the lack of user-friendly software has limited the use of those designs. Thus, there is a pressing need for developing model-assisted designs that can complement the current designs and for building reliable, interactive, and freely available software to facilitate the use of the novel designs. In this work, we first clarify some mis-conceptions between algorithm-based designs and the model-assisted designs with the purpose to eliminate the confusion caused by the designs' similarity in their appearance. Second, we develop a class of novel model-assisted designs that aim to accommodate the urgent need to utilize readily available historical data or real-world evidence to further improve the efficiency of the model-assisted Phase I designs. Third, we construct a seamless early phase trial design that addresses the challenges emerging along with the vast development of immunotherapy and targeted therapy. Fourth, we develop a versatile software platform to provide user-friendly web-based applications to facilitate the use of a series of well-performed model-assisted designs that are built on sound statistical foundations and have su-

perior operating characteristics (e.g., have high probability of identify the MTD and treat a large number of patients at the MTD). In addition to the important issues addressed for model-assisted Phase I designs, we also include the examination of a critical topic in Phase II clinical trials: sequential monitoring. We thoroughly studied the connections between different sequential monitoring approaches theoretically. Furthermore, we conduct extensive simulations to examine the impact of different type of prior distributions on the false positive rate and power to test the efficacy of a treatment, and provide practical recommendations for Phase II sequential monitoring. Our research will greatly advance drug development as it not only provides a wide range of innovative designs, but also creates user-friendly versatile software platforms to facilitate the implementation of the novel designs. In the text below, we provide an overview of each of the five projects.

1.1 Model-assisted design versus algorithm-based design

Comprehensive comparisons among algorithm-based, model-based, and model-assisted designs have been conducted in previous studies [85, 124, 125]. The studies showed that model-assisted designs (e.g., BOIN and keyboard) outperform algorithm-based designs in terms of the accuracy, safety, and reliability, and have comparable performance to more complicated model-based designs. BOIN is a well validated design and has gained a lot of popularity in various types of Phase I clinical trials [114]. The most well-known hallmark of the BOIN design is its concise decision rule: making the decision of dose escalation and de-escalation by simply comparing the observed DLT probability at the current dose with a pair of optimal dose escalation and de-escalation boundaries. Recently, an interval 3+3 (i3+3) design has been proposed and shown to have comparable performance to BOIN in the original paper

[60]. The i3+3 design is an algorithm-based dose-finding design using a BOIN-like decision rule with some modifications. The similarity in the appearance of the two designs has caused substantial confusions among practitioners. In this project, we aim to demystify the i3+3 design by elucidating its links with the BOIN design and compare their similarities and differences, as well as pros and cons. The comparison is illustrated in both their theoretical foundations and operating characteristics based on extensive simulation studies.

1.2 Use informative prior to improve the efficiency of model-assisted Phase I clinical trial designs

Incorporating historical data or real-world evidence has a great potential to improve the success of Phase I clinical trials and accelerate drug development. For model-based Phase I designs such as CRM [77], this can be conveniently carried out by specifying a “skeleton”, i.e., the prior estimate of DLT probability at each dose. In contrast, little work has been done to incorporate historical data or real-world evidence into model-assisted designs, such as BOIN, keyboard, and mTPI designs. This has led to the misconception that model-assisted designs cannot incorporate informative prior information. To improve the efficiency of the model-assisted designs, we propose a unified framework that allows incorporating historical data or real-world evidence into the derivation of the decisions rules in these designs. The proposed approach adapts the well-established “skeleton” approach, combined with the concept of prior effective sample size. It is easy to understand and use. In addition, this approach maintains the hallmark of the model-assisted design: simplicity—the dose escalation/de-escalation rules can be tabulated prior to the trial conduct [114].

1.3 Seamless Phase I/II design for identifying optimal biological dose

In the era of targeted therapy and immunotherapy, the objective of dose finding is to identify the optimal biological dose (OBD) instead of the MTD. The OBD is typically the most desirable dose that considers the trade-off between toxicity and efficacy. Thus, traditional Phase I trials that evaluate toxicity information alone by assuming the monotonicity of dose-toxicity curve may fail to find the optimal dose for subsequent trials. This is because the monotonicity assumption for dose-toxicity relationship implies a dose-response curve that also monotonically increases with dose level. Although it is reasonable to assume that toxicity increases with the dose, the same is not necessarily true for efficacy, since clinical research has shown that the dose-response curve can plateau or even decrease after a certain level of doseage of some immunotherapies or targeted therapies [83]. Thus considering both toxicity and efficacy should be an ideal approach, which indicates the need for seamless Phase I/II designs [118]. There are three general optimization approaches in Phase I/II designs to deal with the trade-off between toxicity and efficacy [117]. The first approach is to set an upper limit on acceptable DLT probability, identify doses with estimated DLT probability less than the limit, and select the optimal dose as the one with the largest efficacy probability; or set a lower limit on acceptable efficacy probability, identify doses with efficacy probability greater than the limit, and select the optimal dose as the one with the smallest DLT probability. The second strategy is to map the joint probability of efficacy and toxicity to a numerical value that quantifies the desirability of the dose, where the numerical value is determined by certain prior specifications and model fitting. Another great alternative is to assign a numerical utility value to all the possible combinations of the toxicity endpoint

and efficacy endpoint, where a model is assumed for all the joint outcomes. The utility quantifies how useful each outcome is for patients using a prespecified scale (e.g., 0-10, or 1-100). With the posterior DLT probability obtained for all doses, the expected posterior mean utilities for the doses are easily determined and used to guide dose assignment in the trial.

One of the classical Phase I/II designs is the EffTox design [97, 96] that uses the second strategy to measure the tradeoff between binary toxicity and efficacy. Yuan and Yin [118] proposed the time-to-event EffTox (TTE-EffTox) that jointly models toxicity and efficacy as time-to-event outcomes. Both of the designs are model-based with certain parametric assumption on the dose-toxicity and dose-response curves. And their decision rules cannot be tabulated. Additionally, quantifying the trade-off based on model fitting is also sensitive to the prior specifications.

To overcome the limitations, we propose a model-assisted Phase I/II design that jointly evaluates toxicity and efficacy using a utility function to measure risk-benefit (i.e., toxicity-efficacy) trade-off after sufficient data are accumulated in stage I using the BOIN design (the design is referred to as U-BOIN). The advantage of the utility-based approach is that it is more straightforward to elicit from clinicians and is also applicable for trials with ordinal or time-to-event outcomes. Specifically, the U-BOIN design consists of two seamlessly connected stages. In stage I, the U-BOIN design quickly explores the dose space and collects preliminary toxicity and efficacy data. In stage II, a multinomial-Dirichlet model is utilized to model toxicity and efficacy endpoints jointly and a utility function elicited from clinicians is employed to measure the dose risk-benefit trade-off. Based on the accumulating efficacy and toxicity from both stages I and II, U-BOIN continuously updates the posterior estimate of the utility for each dose after each cohort, and uses this information to

direct the dose assignment and selection. Compared to existing Phase I/II designs, one prominent advantage of the U-BOIN design is its simplicity for implementation. Once the trial is designed, it can be easily applied using predetermined decision tables, without complex model fitting and parameter estimation. Our simulation study shows that, despite its simplicity, the U-BOIN design is robust and has high accuracy to identify the OBD. In this project, we also extend the design to accommodate delayed efficacy by leveraging the short-term endpoint (e.g., immune activity or other biological activity of targeted agents), and using it to predict the delayed responses to facilitate real-time decision making.

1.4 Versatile software platform of Bayesian adaptive designs for early phase clinical trials

It has been shown in previous studies [124, 125] that the model-assisted designs BOIN and keyboard have greater operating characteristics in that they both perform better than 3+3 in terms of accuracy and reliability, and are safer than model-based designs such as CRM and other model-assisted designs (e.g., mTPI). Moreover, the use of the designs can result in larger success rate in later phases [85]. The standard BOIN and keyboard designs can be used only in single-agent trials with binary endpoint and requires fully observed toxicity outcomes for all patients in the trial in order to make dose assignment decisions. Researchers have extended the two designs to accommodate commonly encountered challenges in Phase I clinical trials, such as fast accrual or delayed toxicity outcomes [116, 57], evaluation of drug combinations [121, 56], identification of the OBD instead of the MTD [126], and incorporation of historical or real-world evidence [127]. All these designs are model-assisted and easy to implement once the decisions rules are determined before trials start. However, the conventional 3+3 still remains as the dominant method

used in Phase I clinical trials to identify the MTD, despite the well-documented flaws of this design. The lack of reliable, robust, and easy-to-use software has limited the use of these novel designs in clinical trials, since it is not convenient for practitioners to evaluate the accuracy and reliability of the designs and they are likely to continue to use the approach they are already familiar with. Lee and Chu [39] suggested that “the availability of accompanying software for the implementation of Bayesian methods is crucial for the use of these methods in clinical trials”. Some software applications have been developed to make model-based designs easier to implement [79, 100, 102]. But few works have been done to make the aforementioned model-assisted designs more accessible for practitioners, even though there are existing R packages, such as BOIN package. The use of a R package would require a certain level of programming knowledge, which is not easy for everyone. Thus, it remains a challenge for practitioners to evaluate the operating characteristics of the designs without programming skills. A broader use of the model-assisted designs requires robust yet user-friendly interactive software.

To facilitate the use of model-assisted designs for designing and conducting Phase I clinical trials and make it more accessible for clinicians who may not have expertise in statistical programming, we develop a web-based software platform for BOIN design and its extensions (BOIN suite), and Keyboard and its extensions (Keyboard suite), respectively. Compared to other existing web-based applications, one distinct feature of our software platform is that each of the two suites includes multiple model-assisted designs that can accommodate different challenges in Phase I clinical trials. The BOIN suite includes designs for single-agent trials using non-informative prior (standard BOIN [65]) or incorporating historical data (iBOIN [127]); for single-agent trials with fast accrual or delayed toxicity (TITE-

BOIN [116]); for drug-combination (BOIN COMB) to find a single MTD [56] or a MTD contour [121]; for finding OBD based on risk-benefit trade-offs in immunotherapy and targeted therapy trials (U-BOIN [126]). The Keyboard suite includes designs for single-agent trials without delayed outcomes (KEYBOARD [105]) or delayed outcomes (TITE-KEYBOARD [57]); and for combination trials to find a single MTD (KEYBOARD COMB [80]). In addition to the interactive and self-explanatory interface of the web-based applications, extensive help files have also been provided to aid the understanding of the designs and navigate the use of the applications.

1.5 Sequential monitoring in Phase II designs

Phase II clinical trials play a critical role in drug development, as their objective is to evaluate the therapeutic effect of a new treatment, and to screen out in-efficacious ones. If the new treatment shows promising treatment effect, further large-scale confirmatory trials can be carried forward. In general, Phase II trials use a short-term and dichotomous endpoint to characterize the patient clinical response to treatment. For example, a binary objective response endpoint can be used to indicate whether a patient has achieved complete or partial response within a predefined treatment course. A Phase II study can be performed using either frequentist or Bayesian designs. Regardless of the type of designs, interim analysis has been the common feature of Phase II clinical trials to prevent from overdosing patients or exposing patients to sub-therapeutic treatment. Most of the frequentist Phase II designs only have two or three stages with the design parameters determined based on numerical searching. When the number of interim analyses exceeds three or when a continuous endpoint is considered, it is computationally infeasible to determine the optimal design parameters for the frequentist designs. Moreover, the frequentist

designs are rigid in the sense that they do not allow any deviation from the original design. For example, for the Simon’s two-stage design [88], the “go/no-go” decisions can only be made at the predetermined interim times, and it will not stop the trial early even when early observed data have already indicated that the drug is indeed futile. To safe guard patients from futile treatments and accelerate the development of efficacious drugs, it is critical to make adaptive decisions based on the accumulated data throughout the trial in a timely fashion such that the trial can be stopped earlier for futility or efficacy.

When more interim analyses are desired, Bayesian methods becomes particularly appealing due to its “we learn as we go” nature [39]. As a result, Bayesian designs usually are more flexible in making multiple interim decisions. Due to the appealing feature of flexibility, recent years have witnessed vast development in Bayesian Phase II clinical trial designs [32, 45, 49, 86, 94, 98, 99, 103, 122]. Overall, existing Bayesian sequential monitoring approaches can be divided into two categories: posterior (or predictive) probability based (PB) [98, 49] or Bayesian hypothesis test based (TB) [45]. For the PB approach, the “go/no-go” decisions are made based on the Bayesian posterior (or predictive) probability that the treatment is futile or effective. If such a probability is greater than a prespecified upper probability cutoff, then the trial can be terminated early, because the experimental treatment is likely to be futile or promising; Otherwise (i.e., this probability is small), there is not adequate information to deliver any conclusion, and the trial continues to collect more data. On the other hand, the TB approach makes the adaptive “go/no-go” decisions based on the Bayesian hypothesis testing framework with the Bayes factor. Although Johnson and Cook [45] argued that the use of Bayes factor in sequential monitoring can gain more efficiency compared to the PB approach and can also elim-

inate a potential source of bias often caused by prior-data conflict, the research on the TB approach for Phase II trials is lacking. Limited examples include the two-stage Bayes factor-based design [20] and the design to identify the maximum effective dose [30]. To get more insight into the TB monitoring, we studied and showed the connection between the TB and PB approaches.

Although the Bayesian approaches gain in popularity in Phase II clinical trials, an issue that has been largely overlooked is the choice of prior distributions. For the TB monitoring approach, a prior distribution is required to quantify the initial uncertainty of the unknown parameter under the alternative hypothesis. The prior distributions for TB monitoring can be classified into local priors and nonlocal priors [46]. A local prior refers to a probability density that has positive density to regions of the parameter space that are consistent with the null hypothesis. In contrast, a nonlocal prior will not overlap with the null hypothesis. The commonly used Beta distribution for the response rate falls into the category of local priors and the inverse moment density (iMOM) is an nonlocal prior. We extensively examined the effects of local and nonlocal priors on the TB monitoring procedure under various trial settings and provide recommendations for practical use. To facilitate sequential TB monitoring, we also develop a web-based software to help clinicians to obtain decision tables and prepare trial protocols.

The five projects introduced above make significant contributions to drug development.

- 1) The thorough comparison between the BOIN and the recently proposed i3+3 provides practitioners the pros and cons of the designs, clarifies some confusions among practitioners, and helps practitioners to choose an appropriate design for early phase trials in order to increase the success of drug development at later

phases.

- 2) The proposed model-assisted designs (e.g., iBOIN) overcomes the limitation of the current model-assisted designs and efficiently make use of the historical data and real-world evidence into the design of Phase I clinical trials.
- 3) The U-BOIN design serves as a novel Phase I/II model-assisted design that uses a utility function to measure dose desirability and can be reliably used for finding the OBD in trials for immunotherapy and targeted therapies.
- 4) The development of versatile software platforms for Phase I Bayesian adaptive designs provides a critical hub for designing and implementing early phase clinical trials. The availability of the free software platforms with regular maintenance allows researchers easily to compare and contrast different designs, and enables clinicians to evaluate the accuracy and safety of the designs easily without knowing any programming knowledge.
- 5) The examination of sequential monitoring in Phase II clinical trials reveals the connection between PB monitoring and TB monitoring and provides practical recommendations for the choice of prior distributions. Along with the user-friendly software, this project complements the existing tools for Phase II trial monitoring.

CHAPTER 2

A comparative study of Bayesian optimal interval (BOIN) design with interval 3+3 (i3+3) design for Phase I oncology dose-finding trials

2.1 Introduction

The objective of Phase I oncology trials is to identify the maximum tolerated dose (MTD) of a new drug, which is defined as the dose with a dose limiting toxicity (DLT) probability that is closest to the target DLT probability. Phase I dose-finding designs can be conducted using algorithm-based designs, model-based designs, or model-assisted designs [36, 48, 115].

Algorithm-based designs determine dose assignment (escalation or de-escalation), based on a set of prespecified rules. Examples include the 3+3 design, accelerated titration design [89], the biased-coin design [21], and its variations [38, 92]. Recently, Liu et al. [60] proposed an algorithm-based design, called interval 3+3 (i3+3), which guides dose escalation and de-escalation by comparing the observed DLT probability with a prespecified toxicity equivalence interval. Algorithm-based designs are simple and easy to implement, but their decision rules are chosen mostly in an ad hoc way and lack of statistical justification. As a result, their statistical properties are often not well defined, leading to poor operating characteristics.

Model-based designs are a class of dose-finding designs that employ a statistical model to describe the dose-toxicity curve and guide dose transition. The most well known model-based design is the continual reassessment method (CRM [77]). The CRM begins with a prior dose-toxicity curve. As information accrues during the trial, it continuously updates the estimate of the model after each cohort, and then uses it to determine the dose assignment for the next cohort of patients. Various extensions of the CRM have been proposed, including dose escalation with overdose control [2], bivariate CRM [5], time-to-event CRM [13], partial order CRM [101], and

Bayesian model averaging CRM [111]. Cheung provides a comprehensive review of the CRM and its related methods [12]. Compared to algorithm-based designs, model-based designs have solid statistical foundation and yield better operating characteristics. However, for appropriate use, the CRM requires specialized expertise to choose and calibrate the dose-toxicity model, and to re-estimate the model at each decision of dose escalation/de-escalation. It remains a challenge to communicate to clinicians how the design works, leading them to perceive the dose allocation rules as coming from a “black box” [67]. The complexity and lack of understanding of the model-based designs result in underutilization of the model-based designs[15, 84].

Model-assisted designs were developed to combine the advantages of algorithm-based designs and model-based designs. Similar to model-based designs, the model-assisted designs use a statistical model (e.g., the binomial model for the number of patients who experience at least one DLT) to derive the decision rules; but like the algorithm-based designs, model-assisted designs can have their dose escalation and de-escalation rules determined before the onset of the trial, and thus can be implemented as simple as the algorithm-based designs. Examples of model-assisted Phase I designs include the modified toxicity probability interval (mTPI) design [42], Bayesian optimal interval (BOIN) design [65, 112], and keyboard design [106] or its equivalent version of mTPI extension (mTPI-2 [31]). Comprehensive reviews and numerical studies have been conducted to compare the model-assisted designs with the 3+3 design and several model-based designs (e.g., CRM), and the results show that the model-assisted designs substantially outperform the 3+3 design, yielding performance comparable to model-based designs [85, 125, 124].

The objective of this chapter is to review and compare the recently proposed algorithm-based i3+3 design with the model-assisted BOIN design. From both theoretical and practical viewpoints, we elucidate their links, similarities, and differences, and pros and cons, supported by extensive simulation studies. The reason we choose to focus on these two designs is that they share considerable similarity in their decision rules, which has caused substantial confusions among practitioners. Our goal is to clarify these confusions, delineate their differences, and provide guidance on choosing appropriate design framework for Phase I clinical trials. We do not include other designs (such as CRM, mTPI and keyboard) in our comparison as comprehensive reviews and comparisons have been provided in Zhou et al. [124, 125] and Ruppert and Shoben [85]. We refer interested readers to these papers.

The remainder of this chapter is organized as follows. In section 2.2, we review the BOIN and i3+3 designs, and delineate their links and difference. In section 2.3, we describe and report

the results of our simulation study. In section 2.4, we conclude with some further discussion of our findings.

2.2 Method

Consider a Phase I trial with J prespecified doses of a drug under examination, $d_1 < \dots < d_J$. Let p_j to denote the DLT probability that corresponds to d_j , and ϕ denote the target DLT probability for the MTD. Let n_j denote the number of patients who have been assigned to d_j , and y_j denote the number of patients who have experienced DLTs at d_j , for $j = 1, \dots, J$. The maximum sample size is N , and patients are enrolled sequentially and treated in cohorts. We assume that the cohort size is prespecified, typically ranging from 1 to 3. Cohort size of 1 leads to a fully sequential trial where the decision of dose escalation and de-escalation is made after each patient is treated. We assume that the DLT can be ascertained quickly such that when a new cohort of patients are ready for enrollment and dose assignment, the DLT outcomes of enrolled patients have been fully observed. In what follows, we first review the BOIN design, followed by the i3+3 design.

2.2.1 Bayesian optimal interval design

Compared to other novel model-based designs (e.g., CRM and BLRM) and model-assisted designs (e.g., mTPI and keyboard), the hallmark of BOIN design is its simple and transparent decision rule: the dose escalation and de-escalation is determined by simply comparing the observed DLT rate $\hat{p}_j = y_j/n_j$ at the current dose with a pair of prespecified, optimized dose escalation boundary λ_e and de-escalation boundary λ_d (see panel (A) in Figure 2.1). Table 2.1 provides the optimal values of λ_e and λ_d for commonly used target DLT probabilities. The BOIN design is described as follows:

1. Treat the first patient or cohort of patients at the lowest dose or a prespecified starting dose.
2. Let j denote the current dose level. Follow the following algorithm to assign a dose to the next patient or cohort of patients.
 - If $\hat{p}_j \leq \lambda_e$, escalate the dose to level $j + 1$;
 - If $\hat{p}_j \geq \lambda_d$, de-escalate the dose to level $j - 1$;
 - Otherwise, i.e., $\lambda_e < \hat{p}_j < \lambda_d$, stay at the current dose level j .
3. Repeat step 2 until the maximum sample size N is exhausted or an prespecified early stopping rule is satisfied. At that point, select the MTD as the dose whose isotonic estimate of the

DLT probability is closest to the target ϕ (note that no MTD is selected if the first dose is eliminated due to safety issues).

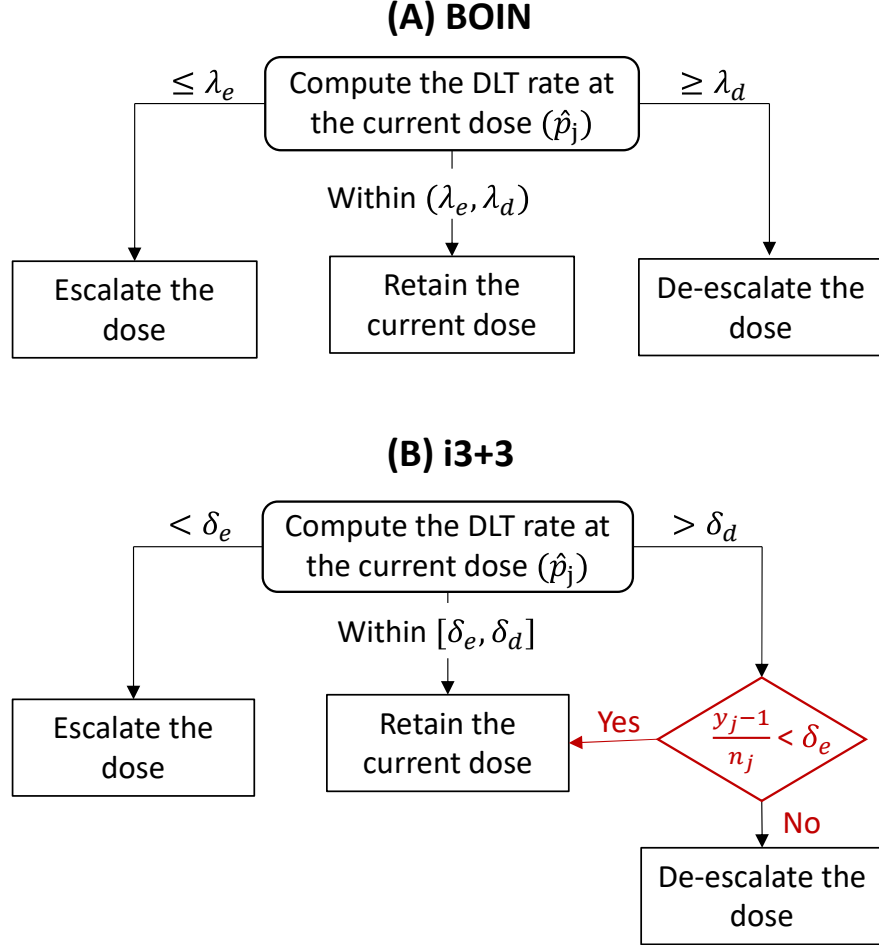


Figure 2.1: Dose escalation and de-escalation algorithm in BOIN and i3+3. DLT rate is the observed DLT probability. The main difference between two designs is that i3+3 adds a “de-escalation modification rule” (in red).

Table 2.1: The optimal escalation/de-escalation boundaries (λ_e, λ_d) under the BOIN design for different target DLT probabilities.

Optimal boundaries	Target DLT probability ϕ					
	0.15	0.20	0.25	0.30	0.35	0.40
λ_e	0.118	0.157	0.197	0.236	0.276	0.316
λ_d	0.179	0.238	0.298	0.359	0.419	0.480

Note. The boundaries (λ_e, λ_d) are calculated using the default under-dosing toxicity probability $\phi_1 = 0.6\phi$ and overdosing toxicity probability $\phi_2 = 1.4\phi$ in Liu and Yuan [65].

The BOIN design is derived based on the optimal design theory. Let $0 < \lambda_e(d_j, n_j, \phi) < 1$ and $0 < \lambda_d(d_j, n_j, \phi) < 1$ denote the generalized dose escalation and de-escalation boundaries that

are unspecified functions of the dose level (i.e., d_j), the number of patients treated (i.e., n_j), and the target DLT probability (i.e., ϕ). The derivation of the BOIN design starts from defining a class of nonparametric designs, denoted as \mathcal{C}_{np} , with the following dose transition rule:

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

Because $\lambda_e(d_j, n_j, \phi)$ and $\lambda_d(d_j, n_j, \phi)$ can freely vary with the dose level d_j , the number of treated patients n_j , and the target ϕ , this class of designs are extremely broad and contain all possible nonparametric designs that do not impose the parametric assumption on the dose-response curve and make dose escalation and de-escalation based on $D_j = (y_j, n_j)$. The mTPI, keyboard/mTPI-2, BOIN, and i3+3 designs all belong to \mathcal{C}_{np} . This provides the first link between i3+3 and BOIN designs. For notation brevity, we use $\lambda_{ej} \equiv \lambda_e(d_j, n_j, \phi)$ and $\lambda_{dj} \equiv \lambda_d(d_j, n_j, \phi)$ hereafter.

The BOIN design is obtained by minimizing the probability of making incorrect decisions of dose escalation and de-escalation within \mathcal{C}_{np} . In other words, the BOIN is the optimal design among \mathcal{C}_{np} with the smallest decision errors. To derive the standard BOIN, Liu and Yuan [65] first define three point hypotheses: $H_0 : p_j = \phi$, $H_1 : p_j = \phi_1$, and $H_2 : p_j = \phi_2$, where ϕ_1 indicates that the dose is substantially under-dosing (i.e., below the MTD) such that escalation is required, and ϕ_2 indicates that the dose is substantially overdosing such that de-escalation is required. Let \mathcal{S} , \mathcal{E} and \mathcal{D} denote stay (at the current dose), escalation, and de-escalation, respectively. Under H_0 , the correct decision is \mathcal{S} , and incorrect decisions are \mathcal{E} and \mathcal{D} ; under H_1 , the correct decision is \mathcal{E} , and incorrect decisions are \mathcal{S} and \mathcal{D} ; and under H_2 , the correct decision is \mathcal{D} , and incorrect decisions are \mathcal{S} and \mathcal{E} . Assuming a non-informative prior that the three hypotheses have equal probability of being true, i.e., $\Pr(H_0) = \Pr(H_1) = \Pr(H_2) = 1/3$, under the dose transition rule of \mathcal{C}_{np} , the probability of making incorrect decision (i.e, decision error) is given by

$$\begin{aligned}
 \alpha &= \Pr(H_0) \Pr(\mathcal{E} \text{ or } \mathcal{D} | H_0) + \Pr(H_1) \Pr(\mathcal{S} \text{ or } \mathcal{D} | H_1) + \Pr(H_2) \Pr(\mathcal{S} \text{ or } \mathcal{E} | H_2) \\
 &= \frac{1}{3} \Pr(\hat{p}_j \leq \lambda_{ej} \text{ or } \hat{p}_j \geq \lambda_{dj} | H_0) + \frac{1}{3} \Pr(\hat{p}_j > \lambda_{ej} | H_1) + \frac{1}{3} \Pr(\hat{p}_j < \lambda_{dj} | H_2) \\
 (2.1) \quad &= \frac{1}{3} \{ \text{Bin}(n_j \lambda_{ej}; n_j, \phi) + 1 - \text{Bin}(n_j \lambda_{dj} - 1; n_j, \phi) \} + \frac{1}{3} \{ 1 - \text{Bin}(n_j \lambda_{ej}; n_j, \phi_1) \} \\
 &\quad + \frac{1}{3} \text{Bin}(n_j \lambda_{dj} - 1; n_j, \phi_2),
 \end{aligned}$$

where y_j is assumed to follow a Binomial model, i.e., $y_j \sim \text{Binomial}(n_j, p_j)$, and $\text{Bin}(b; m, \xi)$ is the cumulative density function of the binomial distribution, with size and probability parameters m and ξ , respectively. Liu and Yuan [65] proved that the optimal escalation and de-escalation boundaries (λ_e, λ_d) that minimize the decision error α are given by,

$$(2.2) \quad \lambda_{ej} = \lambda_e = \frac{\log\left(\frac{1-\phi_1}{1-\phi}\right)}{\log\left\{\frac{\phi(1-\phi_1)}{\phi_1(1-\phi)}\right\}}, \quad \lambda_{dj} = \lambda_d = \frac{\log\left(\frac{1-\phi}{1-\phi_2}\right)}{\log\left\{\frac{\phi_2(1-\phi)}{\phi(1-\phi_2)}\right\}}.$$

Remarkably, the optimal escalation and de-escalation boundaries are independent of d_j and n_j , resulting in BOIN's hallmark conciseness: dose escalation and de-escalation is made by a simple comparison of \hat{p}_j with λ_e and λ_d .

Such simplicity might lead one to think that the BOIN decision rule does not consider the variance of \hat{p}_j (or equivalently the sample size n_j). This, however, is not true. As shown by equation (2.1), in the BOIN design, the derivation and minimization of the decision error α depends on the sampling distribution of \hat{p}_j , thus directly taking the uncertainty of \hat{p}_j into account. Liu and Yuan [65] further proved that λ_e and λ_d are also the boundaries corresponding to the likelihood ratio test and Bayes factor, which certainly account for data variability or equivalently sample size n_j . In other words, n_j being cancelled out should not be confused with ignoring n_j . In contrast, as described later, ignoring the variation of \hat{p}_j is a deficiency of the i3+3 design. The values of ϕ_1 and ϕ_2 can be elicited from physicians and should not be set as too close to ϕ as the small sample size of Phase I trial provides very limited power to distinguish a very small difference. Liu and Yuan [65] recommend the default values $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$, which generally yield a design with good operating characteristics. In some cases, the values of ϕ_1 and ϕ_2 can be calibrated to fit specific trial needs. For example, if a stronger overdose control is needed, we may set $\phi_2 = 1.2\phi$ to encourage easier dose de-escalation.

Similarly, because λ_e and λ_d are independent of n_j , one might concern that λ_e and λ_d will not converge to ϕ asymptotically. This concern overlooks the basic fact that when studying the asymptotic property, it is not meaningful to consider fixed alternative hypotheses, but should consider local hypotheses, where H_1 and H_2 converge to H_0 [50]. Thus, to study the convergence of BOIN, appropriate hypotheses are $H_1 : \phi_1 = \phi - d_1/\sqrt{n_j}$ and $H_2 : \phi_1 = \phi + d_2/\sqrt{n_j}$, where d_1 and d_2 are constants. Under these local hypotheses, λ_e and λ_d (and thus the BOIN design) converge to ϕ . As one of the most important features of Phase I trials is small sample size, the asymptotic property is mainly of the theoretical interest with little practical relevance.

For patient safety, BOIN imposes an overdose control/early stopping rule: if $\Pr(p_j > \phi | n_j, y_j) > C$ and at least 3 patients have been treated, the current dose d_j and higher doses are eliminated from the trial, where C is a probability cutoff, typically set as $C = 0.95$. The trial is terminated if the lowest dose is eliminated. The posterior probability $\Pr(p_j > \phi | n_j, y_j)$ is evaluated based on the beta-binomial model with $y_j | p_j \sim \text{Binomial}(n_j, p_j)$ and $p_j \sim \text{Beta}(1, 1)$.

2.2.2 Interval 3+3 design

The i3+3 design is an algorithm-based design and consisted of a set of prespecified rules. It requires the specification of a MTD equivalence interval (EI), defined as $[\phi - \epsilon_1, \phi + \epsilon_2]$, where ϵ_1 and ϵ_2 are small margins, such as 0.05. EI represents a range of toxicity probability (or dose) that can be viewed as practically equivalent to the target DLT probability ϕ (or MTD). That is, EI represents a proper dosing interval. For ease of exposition, define $\delta_e \equiv \phi - \epsilon_1$ and $\delta_d \equiv \phi + \epsilon_2$. The i3+3 design is consisted of the following rules (also shown in panel (B) of Figure 2.1).

1. Treat the first patient or cohort of patients at the lowest dose or a prespecified starting dose.
2. Let j denote the current dose level. Follow the following procedure to assign a dose to the next patient or cohort of patients.
 - If $\hat{p}_j < \delta_e$, escalate the dose to level $j + 1$;
 - If $\hat{p}_j > \delta_d$, de-escalate the dose to level $j - 1$;
 - **(De-escalation modification rule)**
However, if $(y_j - 1)/n_j < \delta_e$, stay at the current dose level j .
 - Otherwise, i.e., $\delta_e \leq \hat{p}_j \leq \delta_d$, stay at the current dose level j .
3. Repeat step 2 until the maximum sample size N is exhausted or the prespecified early stopping rule is reached. At that point, select the MTD as the dose whose isotonic estimate of the DLT probability is closest to the target ϕ but not greater less than δ_d .

The i3+3 design employs a similar overdose control/early stopping rule as described previously. Suppose that the decision is to escalate the dose from level j to $j + 1$, if $\Pr(p_{j+1} > \phi | n_{j+1}, y_{j+1}) > 0.95$, then dose d_{j+1} and higher doses are eliminated from the trial. The trial is terminated if the lowest dose is eliminated.

2.2.3 Comparison of BOIN and i3+3 designs

Contrasting the decision rule of the i3+3 design with that of the BOIN design (see Figure 2.1), several links, similarities, and differences become clear. First, despite its name, the i3+3 design has very little to do with the conventional 3+3 design. Instead, its decision rule is strikingly akin to the BOIN’s hallmark decision rule—make the decision of dose escalation and de-escalation by comparing \hat{p}_j with a pair of dose escalation and de-escalation boundaries (Figure 2.1). The difference between the two designs is that the dose escalation de-escalation boundaries of the BOIN design (i.e., λ_e and λ_d) are derived based on rigorous statistical theory and optimized to minimize the probability of making incorrect dose assignment decisions, whereas those of the i3+3 design (i.e., δ_e and δ_d) are ad hoc and subjectively picked. Such difference stems from the different frameworks on which the two designs are based. The i3+3 design takes the algorithm-based framework and thus inevitably inherits its deficiency—the decision rules are ad hoc and often arbitrarily chosen without any formal statistical justification to ensure statistical properties. As a result, such type of designs have been widely criticized for the lack of consistence and reproducibility because anyone can arbitrarily add, remove or modify rules without rigorous justification. The i3+3 can have similar desirable performance to BOIN in some cases when $\delta_e = \lambda_e$ and $\delta_d = \lambda_d$, otherwise, it could result in undesirable outcomes due to the subjectivity of the specifications for the EI.

Moreover, the use of statistically unfounded ad hoc rule leads to deeper statistical issues. The cornerstone of the i3+3 design is the definition of EI, which says that when the true DLT probability $p_j \in [\delta_e, \delta_d]$, the drug is dosed properly and can be accepted as the MTD. Therefore, when $p_j < \delta_e$, the drug is under-dosing and thus should be escalated, and when $p_j > \delta_d$, the drug is overdosing and thus should be de-escalated. Note that in this definition, p_j is the *true* DLT probability. After defining the EI, the i3+3 design directly replaces p_j with its point estimate \hat{p}_j to make the decision of dose escalation and de-escalation. Because of ignoring the difference between p_j and \hat{p}_j , the i3+3 design fails to account for the variation of \hat{p}_j . To alleviate this problem, the i3+3 design adds an additional ad hoc rule, i.e., the “de-escalation modification rule”. However, this rule does not really address the problem, but instead causes more problems, as described below. In contrast, the BOIN design rigorously accounts for data variability through its derivation and optimization of the dose escalation and de-escalation boundaries as demonstrated in (2.1). It does not need such an additional rule, making BOIN more concise and transparent than i3+3 (Figure 2.1).

Structure-wise, the main difference between BOIN design and i3+3 design is that the latter

adds a “de-escalation modification rule”: if $\hat{p}_j > \delta_d$ and $(y_j - 1)/n_j < \delta_e$, then rather than de-escalating the dose, the dose stays at the current dose level. Liu et al. [60] claimed that this rule accounts for data variability (i.e, the sampling variance of \hat{p}_j). Unfortunately, this ad hoc rule does not serve its intended purpose, but instead causes a series of logical and scientific flaws. First, it is odd that the data variability is considered only when the dose de-escalation criteria is met, but omitted when the dose escalation or stay criteria is met. This contradicts with the clinical practice that extra caution should be taken when performing dose escalation, not when performing dose de-escalation. Also, it is difficult to understand why using $(y_j - 1)/n_j$, but not others, such as $(y_j - 2)/n_j$, or more naturally the variance of \hat{p}_j , to account for the uncertainty of \hat{p}_j .

Secondly, the “de-escalation modification rule” directly contradicts with the essence of the i3+3 design itself, which says that if the observed DLT rate is greater than the upper limit of EI, the dose is regarded as overdosing and thus should be de-escalated. Suppose the target DLT probability $\phi = 0.1$ with EI=[0.05, 0.15], then when $\hat{p}_j = 1/3 = 33\%$ patients has DLT, the “de-escalation modification rule” will force the design staying at that dose to treat patients even through \hat{p}_j is much greater than 0.15, and more precisely $\Pr(p_j > 0.1 | n_j = 3, y_j = 1) = 94.8\%$ based on the beta-binomial model assuming a uniform prior for p_j .

Thirdly, numerically it can be shown that the “de-escalation modification rule” is rarely triggered. For example, when target $\phi = 0.3$ and EI= [0.25, 0.35], this rule will never be triggered when the commonly cohort size (e.g., 3) is used; and when target $\phi = 0.2$ and EI= [0.15, 0.25], this rule is triggered only when 1/3 patients experience DLT. In other words, this “de-escalation modification rule” is largely redundant, playing little role but unnecessarily complicating the design. Moreover, as shown later in the simulation study, adding the “de-escalation modification rule” actually impairs the performance of the design.

All aforementioned issues highlight the deficiency of the algorithm-based design framework that the 3+3 and i3+3 designs are based on: arbitrarily specifying decision rules without rigorous statistical reasoning and theory. These rules often do not serve the purposes they are intended to, but often cause a series of scientific and logistical flaws.

In contrast to the algorithm-based i3+3 design, the BOIN design takes the model-assisted framework. Its decision rule is rigorously derived based on well established statistical theory. This makes the BOIN design more reliable with well known statistical properties [65]. More importantly, the BOIN is also simpler and has been well validated in practice. It has been used in a variety of oncology trials, including trials for pediatric tumors, adult tumors, solid tumors, and liquid tumors.

A list of specific trials can be found in Yuan et al. [115].

2.3 Simulation study

2.3.1 Simulation configuration

To avoid cherry-picking and provide objective comparison, the dose-toxicity scenarios used for comparing the i3+3 and BOIN designs were randomly generated using the pseudo-uniform scenario algorithm proposed by Clertant and O’Quigley [16]. Given a target toxicity rate ϕ and J dose levels, we generated scenarios as follows:

- 1). Select one of the J dose levels as the MTD with equal probabilities.
- 2). Sample $M \sim \text{Beta}(\max\{J - j, 0.5\}, 1)$, where j denotes the selected dose level.
- 3). Set an upper bound $B = \phi + (1 - \phi) \times M$ for the toxicity probabilities.
- 4). Repeatedly sample J toxicity probabilities uniformly on $[0, B]$ until these correspond to a scenario in which dose level j is the MTD.

To ensure that MTD is uniquely and meaningfully defined, we required that the true DLT probability of the MTD is within $[\phi - 0.05, \phi + 0.05]$, and the distance between the MTD and its adjacent doses is greater than 0.05 and less than 0.3, i.e., $0.05 < p_{j+1} - p_j < 0.3$ and $0.05 < p_j - p_{j-1} < 0.3$.

We considered target DLT probabilities of $\phi = 0.2, 0.3$ or 0.4 , each with $J = 5$ dose levels, and a maximum sample size of $N = 30$ with the cohort size of 1 or 3. Under each setting, we generated 1000 scenarios. Figure A2.1 displays 50 randomly chosen scenarios for each ϕ , which cover a variety of dose-toxicity curve shapes and spacing. For each of the 1000 scenarios, we simulated 1000 trials and calculated the following performance metrics:

- (a) *Percentage of correct selection (PCS)*: the percentage of simulated trials in which the MTD is correctly selected.
- (b) *Patient allocation*: the average number of patients assigned to the MTD.
- (c) *Overdose control*: the percentage of patients assigned to the doses that are above the MTD.

For the i3+3 designs, we set the EI as $\delta_e = \phi - 0.05$ and $\delta_d = \phi + 0.05$, which is reasonable based on its authors’ recommendation. For the BOIN design, we used its default optimal boundaries with $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$.

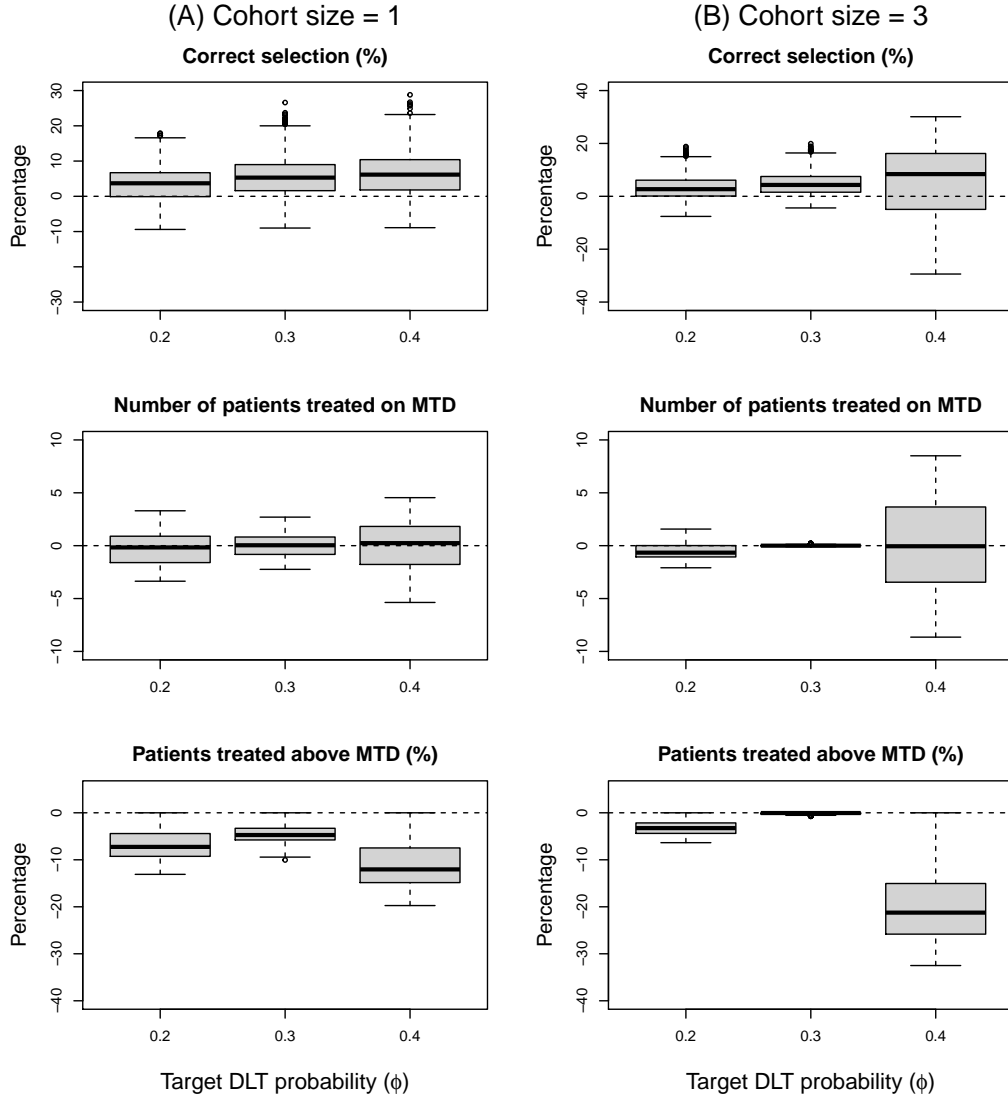


Figure 2.2: Performance difference between BOIN and i3+3 under 1000 randomly generated scenarios. A positive value for “Correct selection (%)” and “Number of patients treated on MTD” indicates that BOIN outperforms i3+3, while a negative value of “Patients treated above MTD (%)” indicates that BOIN is safer.

2.3.2 Results

To summarize the results, we calculate the difference between BOIN and i3+3 in each of the performance metrics, defined as (the performance metric of BOIN) – (the performance metric of i3+3). For metrics (a) and (b), a positive value indicates that BOIN outperforms i3+3, while for metric (c), a negative value indicates that BOIN outperforms i3+3. As 1000 scenarios were considered, there were 1000 values for the difference in each performance metric.

Figure 2.2 shows the results, with each performance metric summarized by boxplots. In

general, BOIN has better accuracy to correctly identify the MTD than i3+3. For example, compared to i3+3, the average percentage of correct selection for BOIN is 5.4% higher when the target DLT probability is 0.3 and the cohort size is 1, and 5.3% higher when the target DLT probability is 0.4 and the cohort size is 3. In addition, BOIN is also safer than i3+3 with better overdose control. For example, i3+3 has 6.4% and 18.3% higher probability of treating patients at the doses above the MTD, when the target DLT probability is 0.2 and cohort size is 1, and when the target DLT probability is 0.4 and cohort size is 3, respectively. On average, the two designs have similar performance in terms of the number of patients treated at the MTD. Figure A2.2 provides the absolute (rather than relative) performance of BOIN and i3+3.

To facilitate a close-up examination on the operating characteristics of the two designs, Table 2.2 shows 15 representative scenarios, 5 for each of the target DLT probabilities ($\phi = 0.2$, 0.3, or 0.4), with the MTD located at different dose levels. Table 2.3 shows the simulation results, which are generally consistent with those from 1000 random scenarios, showing that BOIN has higher accuracy to identify the MTD and better overdose control than i3+3.

Table 2.2: Fifteen representative randomly generated scenarios

Scenario	Dose level				
	1	2	3	4	5
Target DLT probability $\phi = 0.2$					
1	0.23	0.49	0.52	0.53	0.89
2	0.12	0.19	0.27	0.38	0.47
3	0.06	0.15	0.23	0.66	0.71
4	0.04	0.06	0.13	0.20	0.30
5	0.03	0.05	0.08	0.11	0.17
Target DLT probability $\phi = 0.3$					
6	0.31	0.42	0.53	0.54	0.81
7	0.03	0.27	0.46	0.48	0.58
8	0.08	0.11	0.30	0.55	0.66
9	0.05	0.08	0.13	0.30	0.39
10	0.02	0.03	0.06	0.15	0.31
Target DLT probability $\phi = 0.4$					
11	0.36	0.45	0.60	0.63	0.73
12	0.18	0.36	0.50	0.60	0.66
13	0.10	0.15	0.40	0.50	0.52
14	0.16	0.19	0.22	0.37	0.51
15	0.02	0.06	0.10	0.22	0.38

Table 2.3: Operating characteristics of BOIN and i3+3 under the 15 representative scenarios shown in Table 2.2.

Scenario	Percentage of correct selection of MTD		Number of patients treated on MTD		Percentage of patients treated above MTD	
	BOIN	i3+3	BOIN	i3+3	BOIN	i3+3
Cohort size = 1						
1	61.8	63.9	17.2	16.2	26.9	33.9
2	38.0	36.9	8.6	7.9	45.1	54.4
3	50.6	55.0	12.1	14.2	9.9	13.4
4	40.4	45.7	9.5	10.2	24.9	33.0
5	71.2	73.3	15.9	18.5	0	0
6	54.1	50.4	15.2	13.3	41.9	49.6
7	62.6	61.1	13.7	12.8	35.4	40.7
8	68.1	70.9	13.7	14.1	20.8	24.0
9	49.2	49.8	10.6	10.7	28.0	32.9
10	69.3	72.2	16.1	17.5	0	0
11	42.4	36.6	13.9	10.1	48.8	63.2
12	52.0	47.7	12.1	10.3	35.8	49.0
13	54.9	46.9	11.8	9.7	27.8	44.3
14	49.3	53.0	9.6	10.6	21.2	33.4
15	76.1	78.0	16.0	19.1	0	0
Cohort size = 3						
1	64.1	63.2	19.7	19.2	15.8	18.2
2	39.3	38.0	9.9	10.1	22.9	28.2
3	45.3	48.0	8.4	9.8	5.3	6.0
4	37.3	37.5	6.4	7.0	10.8	14.3
5	54.0	62.4	7.6	9.5	0	0
6	57.5	57.4	17.6	17.5	32.5	32.9
7	67.9	67.1	15.1	15.0	27.0	27.2
8	69.9	69.6	12.9	12.9	15.4	15.4
9	48.6	48.8	9.1	9.1	17.1	17.1
10	71.1	71.1	11.5	11.5	0	0
11	55.4	39.8	19.6	11.8	30.4	58.6
12	61.3	46.5	14.4	10.4	19.3	45.0
13	63.8	43.1	12.0	9.2	11.8	37.7
14	46.6	49.6	6.6	8.2	7.1	29.1
15	69.9	77.4	9.1	14.6	0	0

2.3.3 The role of “de-escalation modification rule”

One main difference between i3+3 and BOIN is that the former adds the “de-escalation modification rule” (see Figure 2.1). Liu et al. (2019) claimed that this rule is a main contribution and the key for the performance of the design. To understand the role of that rule, we compared the performance of the i3+3 design (with the rule) to its counterpart without the rule (denoted as the i3+3s design, where “s” means simplified) under the 1000 random scenarios described previously. That is, the two designs are exactly same except that i3+3s removes the “de-escalation modification rule”. For each scenario, we simulated 1000 trials and calculated the difference between i3+3s and i3+3 in each of the performance metrics, defined as (the performance metric of i3+3s) – (the performance metric of i3+3).

Figure 2.3 shows the results. We can see that the “de-escalation modification rule” has virtually no, and sometimes detrimental, effect on the design performance. In terms of the percentage of correct selection, the mean difference between i3+3 and i3+3s is less than 1.6% in all the simulation settings. This is particularly obvious when the target DLT probability is 0.3 and the cohort size is 3. In this case, the two designs are exactly same as the “de-escalation modification rule” is never triggered. The same pattern is observed for the number of patients treated at the MTD, i.e., the mean difference between i3+3s and i3+3 is less than 0.4 (i.e., 1.3%), suggesting that adding the rule provides virtually no gain in this metric either. For safety, adding the rule actually led to substantially worse performance, and resulted in treating 2.7% to 13.4% more patients on the toxic doses above the MTD, except when the target DLT probability is 0.3 and the cohort size is 3 (as noted above, the two designs are equivalent in this case).

These results validate the previous theoretical analysis, showing that the “de-escalation modification rule” is not only redundant, but often detrimental. This exemplifies the risk of arbitrarily adding rules without careful and formal statistical and scientific justification for algorithm-based designs. Removing the “de-escalation modification rule” from i3+3 improves the performance of the design. The dilemma is that the resulting design (i.e., i3+3s) then is a trivial variation of BOIN. From this perspective, i3+3 is simply a downgraded version of BOIN.

2.4 Summary

We have elucidated the similarity and difference between the recently proposed algorithm-based i3+3 design and the model-assisted BOIN design. We showed that the i3+3 design adopted very similar decision rules as the BOIN design, but because of taking the algorithm-based design

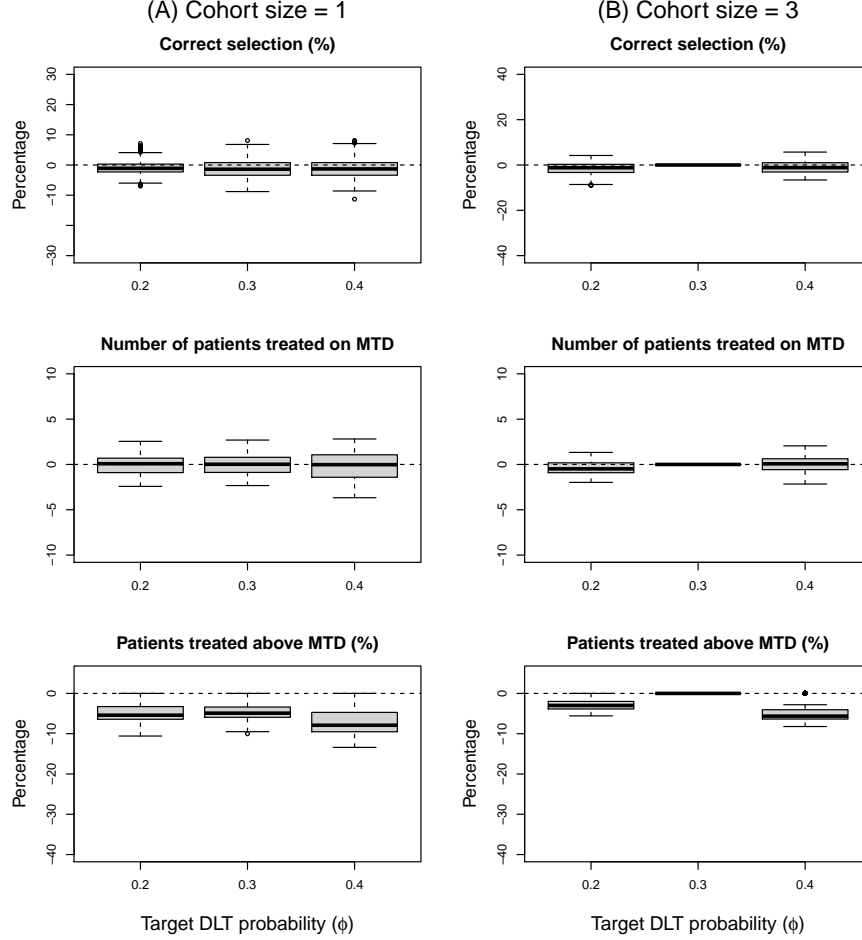


Figure 2.3: Performance difference between i3+3s (i.e., the counterpart of i3+3 with the “de-escalation modification rule” removed) and i3+3 under 1000 random scenarios. A positive value for “Correct selection (%)” and “Number of patients treated on MTD” indicates that i3+3s outperforms i3+3, while a negative value of “Patients treated above MTD (%)” indicates that i3+3s is safer.

framework, it suffers from a series of scientific and logical deficiencies inherited in algorithm-based designs. We further showed that the “de-escalation modification rule” used by i3+3 to differentiate it from BOIN does not improve, but often impair the design performance. Simulation studies show that compared to the i3+3 design, the BOIN design has better accuracy to identify the MTD and is safer with a lower risk of overdosing patients. In addition, BOIN is also simpler, more transparent, and has been widely validated in a variety of clinical trials. Thus, we believe BOIN provides a better choice for Phase I oncology trials.

The solid statistical foundation underlying the BOIN design provides its ability to be extended to handle more complicated clinical applications. Under the model-assisted framework, the BOIN design has been extended to accommodate drug-combination [56], late-onset toxicity

[116], efficacy-toxicity based Phase I/II trials [93, 126]. More details for these extended designs are provided in Chapters 4 and 5.

A Appendix

A2.0.1 A subset of random scenarios when comparing BOIN to i3+3

Figure A2.1 shows a random sample of the 1000 random scenarios used for comparing BOIN and i3+3 design under each target DLT probabilities.

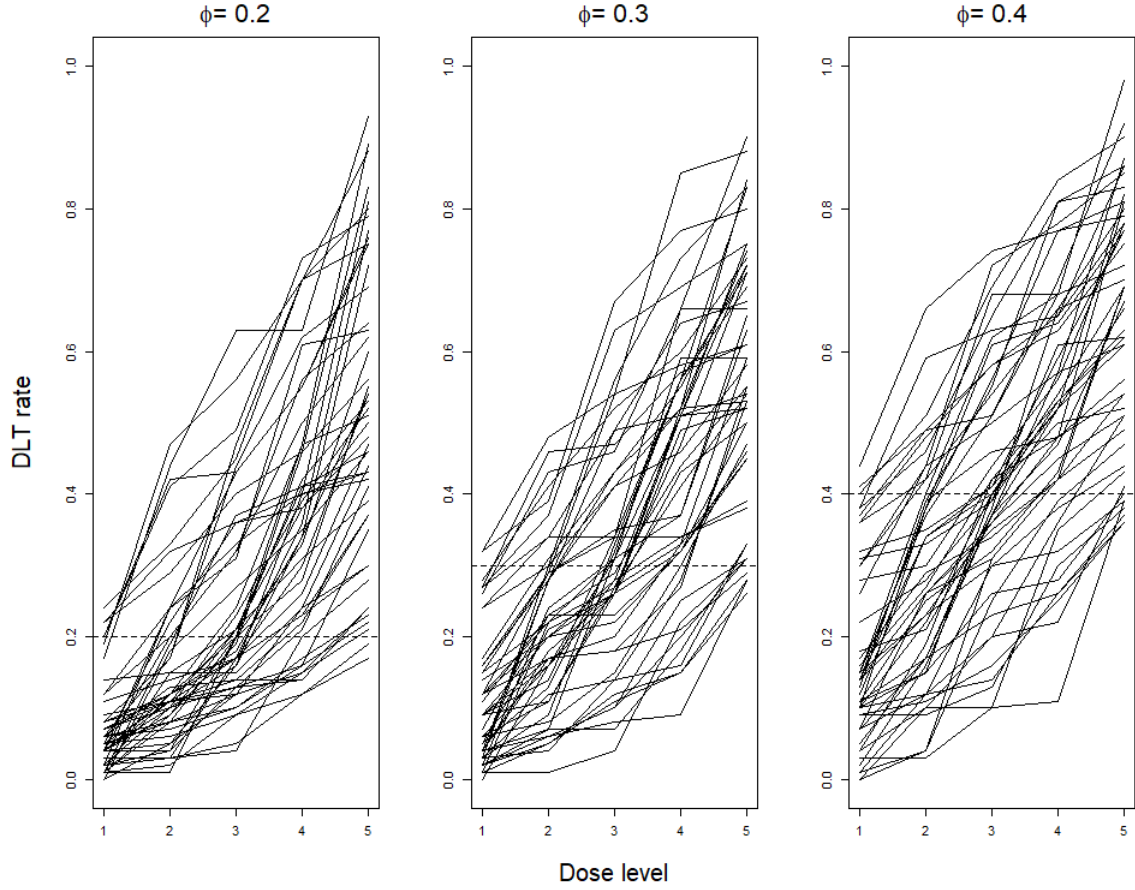


Figure A2.1: A sample of 50 randomly generated scenarios given the target DLT probability $\phi=0.2$, 0.3 or 0.4.

A2.0.2 The simulation results for BOIN and i3+3 designs under 1000 random scenarios

Figure A2.2 shows the simulation results for BOIN and i3+3 designs under 1000 random scenarios, under different target DLT probabilities. Both designs were implemented under their default/recommended settings. As shown, BOIN has larger probability of correctly identifying the MTD and is safer in that it assigns less patients on doses above MTD. The number of patients treated on MTD is comparable between the two designs, but i3+3 is still more aggressive than BOIN and assigns more patients on doses above MTD, an undesirable consequence caused by i3+3's "de-escalation modification rule".

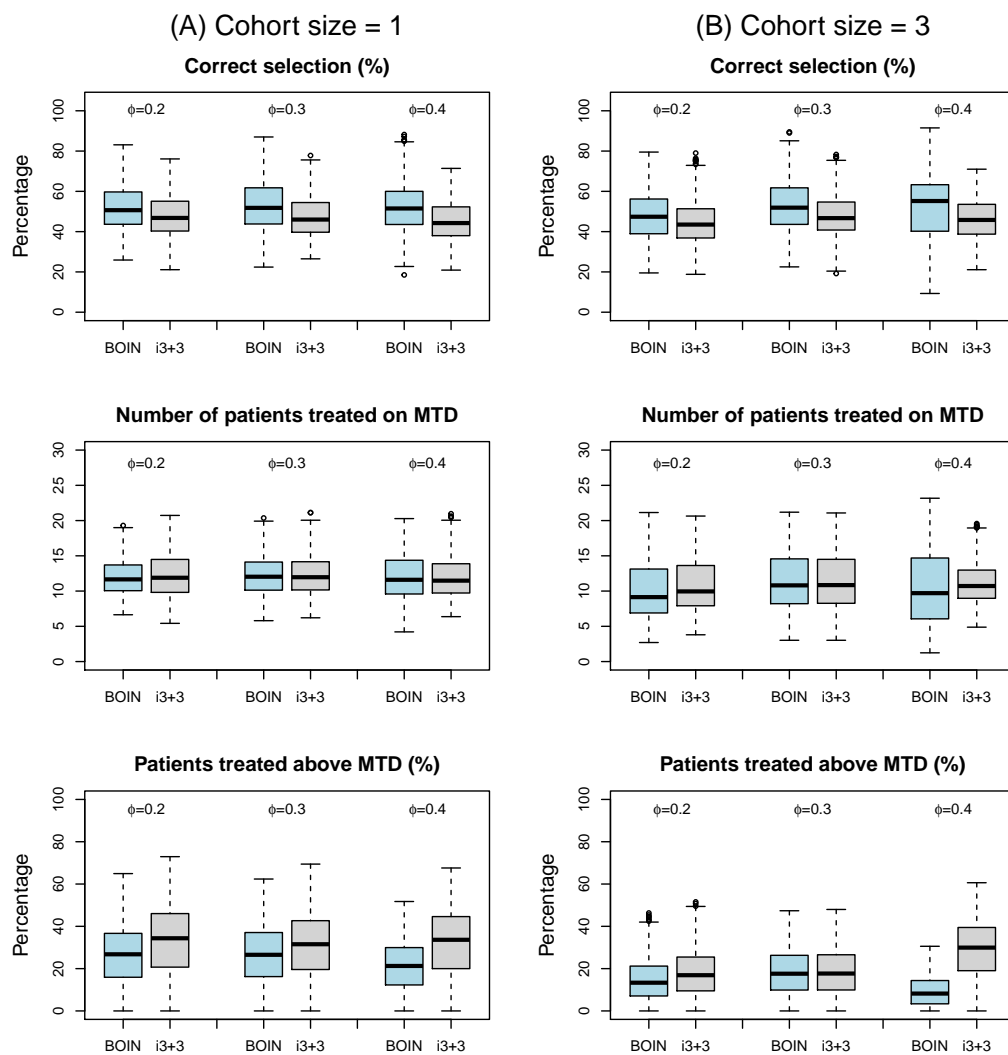


Figure A2.2: Performance of BOIN and i3+3 designs under 1000 randomly generated scenarios given different target DLT probability (ϕ) and cohort size.

CHAPTER 3

Incorporating historical information to improve phase I clinical trial designs

3.1 Introduction

Recently, there is tremendous interest to use prior information, such as historical data or real-world evidence, as an effective way to improve the efficiency of clinical trials. In May 2019, Food and Drug Administration (FDA) released a draft guidance on submitting documents using real-world data or evidence to FDA for drugs and biologics [24]. When designing phase I trials, prior information is often available from previous studies. For example, the drug to be investigated has been studied previously in other indications, or similar drugs belonging to the same class have been studied in earlier phase I trials [128]. Another example is the proposal of bridging phase I trials to extend a drug from one ethnic group (e.g., Caucasian) to another (e.g., Asian) [63], or from adult patients to pediatric patients [81]. In this case, dose-toxicity data from the original trial in one ethnic group or adult can be used to inform the design of the subsequent bridging trials, see for example Morita [68], Liu et al. [63], and Li and Yuan [53].

Various Phase I trial designs have been proposed to find the the maximum tolerated dose (MTD). These designs can be classified into algorithm-based, model-based, and model-assisted designs, depending on their statistical foundations and implementation approaches. Algorithm-based designs, such as the 3+3 design, are ad-hoc, simple to implement, but rigid and have poor accuracy to identify the MTD. It is difficult, if not impossible, to incorporate prior information into algorithm-based designs. Model-based designs assume a dose-toxicity model and determine the dose escalation/de-escalation by continuously updating the estimate of the model based on accrued data. Typical examples of the model-based design are the continual reassessment method (CRM [77])

and its variations, such as escalation with overdose control (EWOC[2]), Bayesian logistic regression model (BLRM [74]), and Bayesian model averaging CRM (BMA-CRM [111]). Model-based designs yield better performance than the 3+3 design in identifying the MTD and allocating more patients to the MTD. Another important strength of model-based designs is that it is straightforward to incorporate prior information. In particular, for the CRM, the prior information can be easily incorporated by specifying a “skeleton” of the dose-toxicity model—the prior estimate of dose limiting toxicity (DLT) probability for each dose. More details are provided later. Along that line, Morita [68] proposed to incorporate informative prior to the CRM; Liu et al. [63] proposed to bridge CRM for phase I clinical trials in different ethnic populations based on Bayesian model averaging; Li and Yuan [53] proposed the continuous reassessment method for pediatric phase I oncology trials (PA-CRM) to leverage trial information from adult trials to pediatric trials.

Model-assisted designs were developed to combine the simplicity of algorithm-based designs with the superior performance of model-based designs. Similar to model-based designs, model-assisted designs use a statistical model (e.g., the binomial model) to derive decision rules for efficient decision making; like algorithm-based designs, model-assisted designs can have their dose escalation and de-escalation rules determined before the onset of a trial, and thus can be implemented as simple as algorithm-based designs. Examples of model-assisted designs include the Bayesian optimal interval (BOIN) design [65], modified toxicity probability interval design (mTPI [42]), and keyboard design [105] (or mTPI-2 [31]). Extensive numerical studies show that the model-assisted designs yield superior performance comparable to more complicated model-based designs, and are increasingly used in practice [115].

Model-assisted designs were developed assuming non-informative prior. Little research has been done on how to incorporate informative prior information into the derivation of these designs. This has led to the misconception that model-assisted designs cannot incorporate informative prior information, which is sometimes cited as their weakness, compared to model-based designs.

In this chapter, we propose a unified framework to incorporate informative prior information into model-assisted designs, including BOIN and keyboard/mTPI-2 designs. Our method utilizes the skeleton approach same as in the CRM, combined with the concept of prior effective sample size (PESS) [69]. The method is intuitive and easy to understand, and more importantly, maintains the simplicity of the model-assisted designs in the sense that their dose escalation/de-escalation rule can still be determined and included in the protocol before the onset of a trial. Numerical studies show that incorporating appropriate informative prior information can improve the performance of

the model-assisted designs in a similar way as that of the CRM.

The remainder of the paper is organized as follows. In section 3.2, we propose the methodology of incorporating informative prior information through skeleton and PESS for CRM, BOIN and keyboard designs. In section 3.3, we provide the software to implement the proposed designs. In section 3.4, we conduct extensive simulation to evaluate operating characteristics of the proposed methodology. We conclude the study with a brief summary in section 3.5.

3.2 Method

3.2.1 Incorporate prior information in CRM

We first describe how to incorporate prior information in the CRM. Let $j = 1, \dots, J$, denote the J doses under investigation, and p_j denote the true DLT probability of dose j . The objective of the trial is to find the MTD, whose DLT probability is equal or the closest to a prespecified target DLT probability ϕ .

To incorporate prior information on the dose-toxicity relationship, we elicit the prior estimate of (p_1, \dots, p_J) , denoted as (q_1, \dots, q_J) , known as “skeleton”. The skeleton should be specified by clinicians based on their clinical experience or estimated using historical data or real-world evidence. We link p_j with the skeleton through a parametric model

$$(3.1) \quad p_j = q_j^{\exp(\alpha)}, \quad \text{for } j = 1, \dots, J,$$

where α is an unknown parameter, which controls the discrepancy between the prior estimate q_j and the true DLT probability p_j . Under the Bayesian paradigm, we assign α a normal prior $f(\alpha) = N(0, \sigma^2)$, where σ^2 is a prespecified hyperparameter. As a result, *a priori* the dose-toxicity curve (p_1, \dots, p_J) centers around the skeleton (q_1, \dots, q_J) . The value of σ^2 controls how much information contained in the prior that can be borrowed from historical data (i.e., the skeleton). A smaller value leads to stronger borrowing. If $\sigma^2 = 0$, the prior completely dominates the observed data and $p_j \equiv q_j$ regardless of the observed data.

Let $D = (D_1, \dots, D_J)$ denote the observed data, where $D_j = (n_j, y_j)$ denotes the data observed at dose level j with n_j being the number of patients treated and y_j the number of patients who experienced DLTs at dose j . To make the decision of dose escalation and de-escalation, the

CRM updates the posterior estimate of p_j as follows,

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

where $L(D | \alpha) = \prod_{j=1}^J \left\{ q_j^{\exp(\alpha)} \right\}^{y_j} \left\{ 1 - q_j^{\exp(\alpha)} \right\}^{n_j - y_j}$ is the likelihood function, and $f(\alpha) = N(0, \sigma^2)$ is the prior distribution of α . Then, the CRM assigns the next cohort of patients at the dose whose \hat{p}_j is closet to ϕ . In practice, we typically impose safety rules, such as starting at the lowest dose level and no dose skipping during dose escalation.

Given a specific skeleton and prior $f(\alpha) = N(0, \sigma^2)$, it is of great importance to quantify how much information is borrowed from historical data. This is, however, rarely discussed in the dose-finding literature. In what follows, we propose a simple and intuitive approach to formally quantify the information borrowed through the skeleton using the concept of “prior effective sample size (PESS)”, which represents the sample size that the prior information is equivalent to. Morita et al. [69] proposed a general methodology to determine PESS, but requires complicated derivation and intensive simulation.

Our approach is simpler and more intuitive, and built upon the following observation: assuming y_j follows a binomial distribution $\text{Binom}(n_j, p_j)$, if p_j follows a beta prior distribution $\text{Beta}(a, b)$, then $a + b$ can be interpreted as the PESS. Our strategy is to approximate the prior distribution of p_j , induced by model (3.1) and $f(\alpha)$, with a beta distribution by matching the first and second moments. Therefore, the PESS can be easily determined. Specifically, given skeleton (q_1, \dots, q_J) and prior $f(\alpha)$, let μ_j and τ_j^2 denote the prior mean and variance of p_j , respectively, with

$$\mu_j = \int p_j f(p_j) dp_j, \quad \tau_j^2 = \int p_j^2 f(p_j) dp_j - \mu_j^2,$$

where $f(p_j)$ is the prior distribution of p_j induced by the prior distribution of $f(\alpha) = N(0, \sigma^2)$, given by

$$f(p_j) = -\frac{1}{2\pi} \exp \left\{ -\frac{\left[\log \left(\frac{\log(p_j)}{\log(q_j)} \right) \right]^2}{2\sigma^2} \right\} \frac{1}{p_j \log(p_j)}.$$

Matching the first and second moments of p_j by a beta distribution $\text{Beta}(a_j, b_j)$, we obtain that

the PESS of the skeleton is $a_j + b_j$, where

$$(3.2) \quad a_j = \frac{\mu_j^2(1 - \mu_j)}{\tau_j^2} - \mu_j, \quad b_j = \frac{a(1 - \mu_j)}{\mu_j}.$$

This reveals a property of the CRM that is barely discussed in the literature but of great implication in practice. Because p_j is a non-linear function of α , once prior $f(\alpha)$ is specified, the PESS for each dose is automatically determined. For example, given skeleton $(q_1, \dots, q_5) = (0.10, 0.19, 0.30, 0.42, 0.54)$ and prior $f(\alpha) = N(0, 0.72)$, the PESS is $(3, 3, 3, 3.1, 3.4)$ for the five doses. As a result, the CRM does not allow users to specify dose-specific prior information or PESS. However, in practice, we often have unequal amount of prior information for different doses. For example, we often have more data at the doses that are below and around the MTD from historical phase I trials. In this case, it is highly desirable to be able to specify different PESS for different doses according to the historical data.

3.2.2 Incorporate prior information in BOIN

We now discuss how to use the skeleton, coupled with the PESS, to incorporate prior information into model-assisted designs such as the BOIN design. To do so, we first briefly describe the genesis of the BOIN design, which lays the ground for the proposed approach. Consider a class of nonparametric designs \mathcal{C}_{np} as follows.

- (a) Patients in the first cohort are treated at the lowest or a prespecified starting dose level.
- (b) At the current dose level j , let $\hat{p}_j = y_j/n_j$ denote the observed DLT probability, and $\lambda_e(j, n_j, \phi)$ and $\lambda_d(j, n_j, \phi)$ denote arbitrary functions of j , n_j and ϕ , serving as the dose escalation and de-escalation boundaries, respectively, with $0 \leq \lambda_e(j, n_j, \phi) < \lambda_d(j, n_j, \phi) \leq 1$. Use the following procedure to assign a dose to the next cohort of patients.
 - Escalate the dose level to $j + 1$, if $\hat{p}_j < \lambda_e(j, n_j, \phi)$;
 - De-escalate the dose level to $j - 1$, if $\hat{p}_j > \lambda_d(j, n_j, \phi)$;
 - Stay the same dose level, j , if $\lambda_e(j, n_j, \phi) \leq \hat{p}_j \leq \lambda_d(j, n_j, \phi)$.
- (c) This process is continued until the maximum sample size is reached.

Note that $\lambda_e(j, n_j, \phi)$ and $\lambda_d(j, n_j, \phi)$ can vary with dose level j , the number of patients treated n_j , and the target ϕ . This class of nonparametric designs includes all possible designs that do not

impose a parametric assumption on the dose-toxicity curve. For notational brevity, in what follows, we suppress arguments in $\lambda_e(j, n_j, \phi)$ and $\lambda_d(j, n_j, \phi)$ and denoted them as λ_e and λ_d .

The BOIN design is obtained by choosing the optimal dose escalation and de-escalation boundaries λ_e and λ_d to minimize the probability of making incorrect dose escalation and deescalation decisions. The optimization is carried out under three point hypotheses:

$$H_1 : p_j = \phi; \quad H_2 : p_j = \phi_1; \quad H_3 : p_j = \phi_2,$$

where ϕ_1 denotes the DLT probability that is deemed substantially lower than the target (i.e., underdosing) such that dose escalation should be made, and ϕ_2 denotes the DLT probability that is deemed substantially higher than the target (i.e., overdosing) such that dose de-escalation is required. Thus, the correct decision under H_1 , H_2 and H_3 is stay, escalation, and de-escalation, respectively; and other decisions are incorrect. For example, under H_1 , escalation or de-escalation are incorrect decisions. Liu and Yuan (2015) showed that optimal dose escalation and deescalation boundaries that minimize incorrect decisions are given by

$$(3.3) \quad \begin{aligned} \lambda_e &= \max \left\{ 0, \frac{\log \left(\frac{1-\phi_1}{1-\phi} \right) + n_j^{-1} \log \left(\frac{\pi_{2j}}{\pi_{1j}} \right)}{\log \left\{ \frac{\phi(1-\phi_1)}{\phi_1(1-\phi)} \right\}} \right\}, \\ \lambda_d &= \min \left\{ 1, \frac{\log \left(\frac{1-\phi}{1-\phi_2} \right) + n_j^{-1} \log \left(\frac{\pi_{1j}}{\pi_{3j}} \right)}{\log \left\{ \frac{\phi_2(1-\phi)}{\phi(1-\phi_2)} \right\}} \right\}, \end{aligned}$$

where $\pi_{kj} = \Pr(H_k)$ is the prior probability that the hypothesis H_k is true at dose level j , where $k = 1, 2, 3$. As a result, the BOIN is the optimal design with the lowest decision errors among all nonparametric designs. Liu and Yuan (2015) recommended default values $\phi_1 = 0.6\phi$ $\phi_2 = 1.4\phi$ that lead to desirable operating characteristics and the decision rule that fits most clinical practice.

When there is no reliable prior information available, we can take the non-informative prior approach and assign the equal probability to each of the three hypotheses being true, i.e., $\pi_{1j} = \pi_{2j} = \pi_{3j} = 1/3$. Then, the optimal boundaries (3.3) become

$$(3.4) \quad \lambda_e^* = \frac{\log \left(\frac{1-\phi_1}{1-\phi} \right)}{\log \left\{ \frac{\phi(1-\phi_1)}{\phi_1(1-\phi)} \right\}} \quad \text{and} \quad \lambda_d^* = \frac{\log \left(\frac{1-\phi}{1-\phi_2} \right)}{\log \left\{ \frac{\phi_2(1-\phi)}{\phi(1-\phi_2)} \right\}},$$

which have the desirable feature that they are independent to the dose level j and the number of patients treated n_j . This means that the same pair of dose escalation and de-escalation boundaries

(λ_d, λ_e) can be used throughout of the trial to make the decision of dose escalation and de-escalation, making the BOIN design particularly simple to implement. That is, if $\hat{p}_j < \lambda_e^*$, escalate the dose; if $\hat{p}_j > \lambda_d^*$, de-escalate the dose; otherwise, stay at the current dose.

When prior information is available, we propose the following procedure to incorporate it into the design:

1. Elicit skeleton (q_1, \dots, q_J) and corresponding PESS (n_{01}, \dots, n_{0J}) , where n_{0j} is the desirable PESS for dose level j , $j = 1, \dots, J$.
2. Determine informative prior for H_k , i.e., π_{kj} , as

$$(3.5) \quad \pi_{kj} = \sum_{x=0}^{n_0} \frac{\phi_k^x (1 - \phi_k)^{n_0-x}}{\sum_{k'=1}^3 \phi_{k'}^x (1 - \phi_{k'})^{n_0-x}} \binom{n_0}{x} q_j^x (1 - q_j)^{n_0-x}.$$

3. Make decision of dose escalation and de-escalation according to the boundaries given in (3.3) with π_{kj} determined in step 2.

The derivation of π_{kj} in step 2 is provided in section A3.0.1. We refer to the resulting design (with informative prior) as iBOIN.

Because of the incorporation of the prior information, the escalation and de-escalation boundaries λ_e and λ_d of iBOIN depend on the dose level j , as well as n_j . Figure 3.1 contrasts the boundaries under non-informative prior and those under informative prior for a trial with 5 doses and the elicited skeleton (0.10, 0.19, 0.30, 0.42, 0.54) when PESS is 3 and 5, respectively. For example, because the prior information says that the lowest dose is under the MTD (with the prior toxicity probability of 0.1), its escalation boundary λ_e thus is lower than that of the non-informative prior to encourage dose escalation. On the contrary, because the prior information says the highest dose is above the MTD (with the prior toxicity probability of 0.54), its de-escalation boundary λ_d thus is lower than that of the non-informative prior to encourage dose de-escalation. When $(n_{01}, \dots, n_{0J}) = 0$, iBOIN becomes the standard BOIN (with non-informative prior).

Compared to the CRM, iBOIN is more flexible and allows users to accurately incorporate prior information by specifying a PESS for each of the dose. For example, given a phase I trial with 5 doses, if historical data provide more information on the first 2 doses than the last 2 doses and most information on dose level 3, we could specify the PESS for 5 doses as (3, 3, 6, 1, 1) to reflect that. As described previously, this is very difficult, if not impossible, under the CRM.

The other advantage of the iBOIN is that its the dose escalation and de-escalation rule can

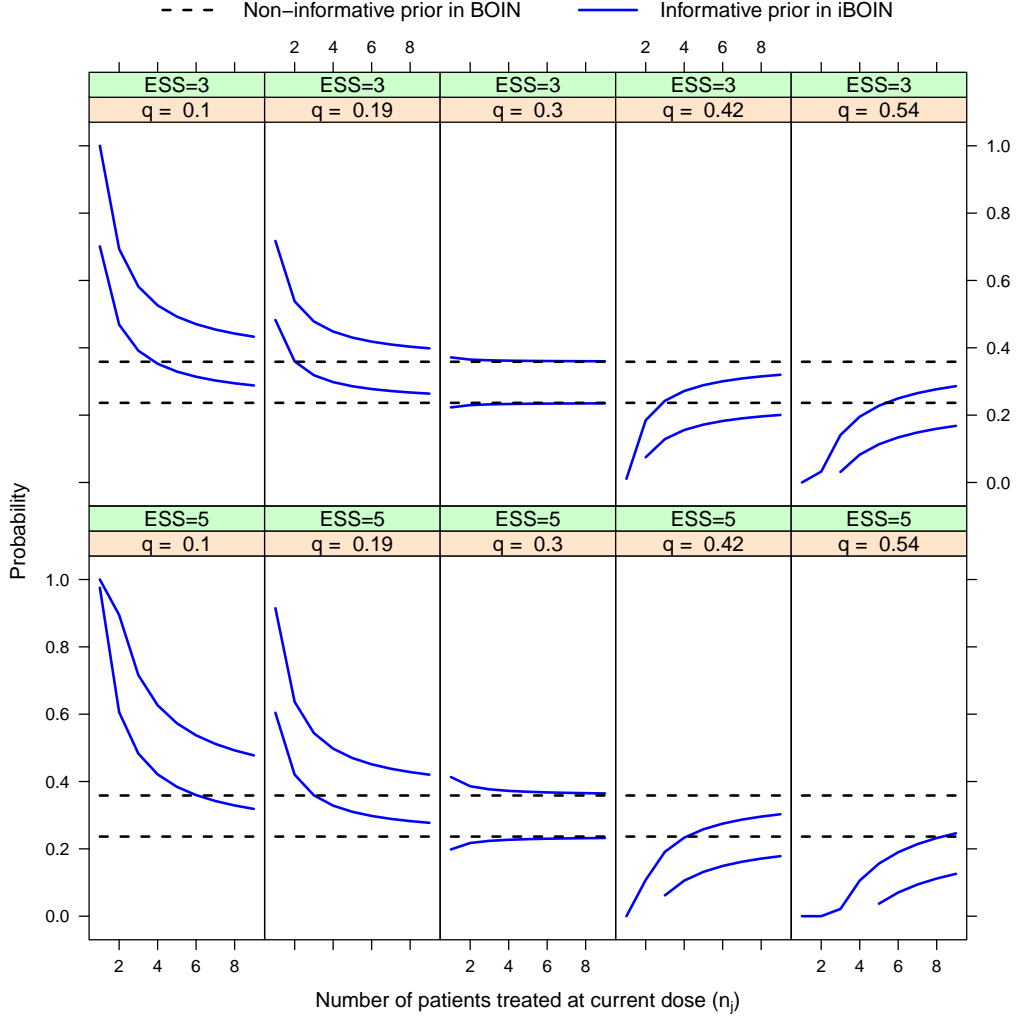


Figure 3.1: Escalation and de-escalation boundaries (λ_e, λ_d) of iBOIN given different prior DLT probability (q) and PESS = 3 or 5, in comparison to the boundaries determined using non-informative prior in standard BOIN.

be pre-tabulated and included in the trial protocol. Table 1 shows the decision table of iBOIN with skeleton (0.10, 0.19, 0.30, 0.42, 0.54) and the effective sample size $n_{01} = \dots = n_{05} = 3$.

This decision table is equivalent to the rule based on λ_e and λ_d , but easier to use in practice. The user only need to identify the row corresponding to the current dose level, and then can use the boundaries listed in that row to easily make the decision of dose escalation and de-escalation. In summary, the iBOIN design 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. Given data (n_j, y_j) observed at the current dose level j , make the decision of escalation/de-

Table 3.1: iBOIN decision boundaries up to 30 patients with a cohort size of 3, given the skeleton $(q_1, \dots, q_5) = (0.10, 0.19, 0.30, 0.42, 0.54)$ and PESS $n_{01} = \dots = n_{05} = 3$. The target DLT probability $\phi = 0.3$.

Dose level	Action*	Number of patients treated at current dose									
		3	6	9	12	15	18	21	24	27	30
1	Escalate if no. of DLT \leq	1	1	2	3	4	4	5	6	6	7
	De-escalate if no. of DLT \geq	2	3	4	5	7	8	9	10	11	12
2	Escalate if no. of DLT \leq	0	1	2	3	3	4	5	5	6	7
	De-escalate if no. of DLT \geq	2	3	4	5	6	7	8	9	11	12
3	Escalate if no. of DLT \leq	0	1	2	2	3	4	4	5	6	7
	De-escalate if no. of DLT \geq	2	3	4	5	6	7	8	9	10	11
4	Escalate if no. of DLT \leq	0	1	1	2	3	3	4	5	6	6
	De-escalate if no. of DLT \geq	1	2	3	4	6	7	8	9	10	11
5	Escalate if no. of DLT \leq	0	0	1	2	2	3	4	5	5	6
	De-escalate if no. of DLT \geq	1	2	3	4	5	6	7	8	10	11

*When neither “Escalate” nor “De-escalate” is triggered, stay at the current dose for treating the next cohort of patients.

escalation according to the iBOIN decision table (e.g., Table 3.1) for treating the next cohort of patients.

3. Repeat step 2 until the prespecified maximum sample size is reached, and then select the MTD as the dose whose isotonicly transformed estimate of p_j is closest to ϕ .

For the purpose of overdose control, following BOIN, the iBOIN design imposes a dose elimination rule: if $\Pr(p_j > \phi | y_j, n_j) > 0.95$ and $n_j \geq 3$, dose level j and higher are eliminated from the trial, and the trial is terminated if the lowest dose level is eliminated, where $\Pr(p_j > \phi | n_j, y_j)$ is evaluated based on the beta-binomial model with uniform prior. As the objective of the dose elimination rule is to protect patients from excessively toxic doses, it is sensible to use uniform prior to evaluate this rule to avoid potential bias due to the misspecification of the prior.

At the end of the trial, iBOIN uses isotonic estimate of p_j to select the MTD (i.e., Step 3). As determining dose escalation/de-escalation and selecting the MTD are two independent components, when the trial is completed, other methods can also be used to determine the MTD. For example, when desirable, we can fit a dose-toxicity model (e.g., a logistic model) as the CRM to select the MTD.

3.2.3 Incorporate prior information in keyboard/mTPI-2 design

Keyboard design is another model-assisted design, developed to address the overdosing issue of mTPI design. Guo et al. [31] proposed a modification of mTPI, known as mTPI-2, which is statistically equivalent to the keyboard design but less transparent and relies on a perplexing

statistical concept and method (e.g., Occam’s razor and model selection). Thus, we here only present the keyboard design. The methodology and simulation results described below are directly applicable to the mTPI/mTPI-2.

The keyboard design starts by specifying a proper dosing interval $\mathcal{I}^* = (\delta_1, \delta_2)$, referred to as the “target key,” and then populates this interval toward both sides of the target key, forming a series of keys of equal width that span the range of 0 to 1. For example, given a target rate of $\phi = 0.30$, the proper dosing interval or target key may be defined as $(0.25, 0.35)$, then on its left side, we form 2 keys of width 0.1, i.e., $(0.15, 0.25)$ and $(0.05, 0.15)$; and on its right side, we form 6 keys of width 0.1, i.e., $(0.35, 0.45)$, $(0.45, 0.55)$, $(0.55, 0.65)$, $(0.65, 0.75)$, $(0.75, 0.85)$ and $(0.85, 0.95)$. We denote the resulting intervals/keys as $\mathcal{I}_1, \dots, \mathcal{I}_K$.

The keyboard design assumes a beta-binomial model,

$$\begin{aligned} y_j | n_j, p_j &\sim \text{Binom}(n_j, p_j) \\ p_j &\sim \text{Beta}(a_j, b_j), \end{aligned} \quad (3.6)$$

where a_j and b_j are hyperparameters. The posterior distribution of p_j arises as,

$$p_j | D_j \sim \text{Beta}(y_j + a_j, n_j - y_j + b_j), \text{ for } j = 1, \dots, J. \quad (3.7)$$

By default, the keyboard design set $a_j = b_j = 1$ to obtain a uniform prior. To make the decision of dose escalation and de-escalation, given the observed data $D_j = (n_j, y_j)$ at the current dose level j , the keyboard design identifies the interval \mathcal{I}_{\max} that has the largest posterior probability, i.e.,

$$\mathcal{I}_{\max} = \operatorname{argmax}_{\mathcal{I}_1, \dots, \mathcal{I}_K} \{\Pr(p_j \in \mathcal{I}_k | D_j); k = 1, \dots, K\}.$$

\mathcal{I}_{\max} represents the interval that the true value of p_j is most likely located, referred to as the “strongest” key by Yan et al. [105]. Suppose j is the current dose level. The keyboard design determines the next dose as follows.

- Escalate the dose to level $j + 1$, if the strongest key is on the left side of the target key;
- Stay at the current dose level j , if the strongest key is the target key;
- De-escalate the dose to level $j - 1$, if the strongest key is on the right side of the target key.

The trial continues until the prespecified sample size is exhausted, and the MTD is selected based

on isotonic estimates of p_j . During the trial conduct, the keyboard design imposes the same dose elimination/early stopping rule as the BOIN design.

As in the beta-binomial model (3.6), $a_j + b_j$ can be interpreted as the PESS. We propose the following procedure to incorporate prior information into the keyboard design:

1. Elicit skeleton (q_1, \dots, q_J) and corresponding PESS (n_{01}, \dots, n_{0J}) , where n_{0j} is the desirable PESS for dose level j , $j = 1, \dots, J$.
2. Determine hyperparameter a_j and b_j in the beta prior (3.6) as follows:

$$(3.8) \quad a_j = n_{0j}q_j; \quad b_j = n_{0j}(1 - q_j), \quad j = 1, \dots, J$$

3. Make dose escalation and de-escalation based on the resulting posterior given by equation (3.7).

We refer the keyboard design with informative prior as iKeyboard design. Given a fixed maximum sample size, all possible outcomes $D_j = (n_j, y_j)$ can be enumerated, and for each possible outcome, the posterior distribution $f(p_j|D_j)$ can be calculated. Therefore, the dose escalation/de-escalation rule of the iKeyboard can be tabulated. The decision table for the ikeyboard design can also be pre-tabulated as that for iBOIN. The above approach is directly applicable to the mTPI design for incorporating prior information. As the keyboard/mTPI-2 design outperforms the mTPI design in both safety and accuracy (Yan et al., [105]; Zhou et al. [124]; Guo et al. [31]), we will not discuss the mTPI.

3.2.4 Robust prior

The performance of aforementioned designs is affected by whether the informative prior is correctly specified. When the informative prior correctly specified, it improves the accuracy of identifying the MTD. However, when the informative prior is misspecified and conflicts with the true dose-toxicity relationship, it may compromise the accuracy of identifying the MTD. In the numerical study described later, we found that when prior is severely misspecified, the prior that sets the MTD at higher doses has more impact on the design performance than the prior that sets the MTD at lower doses, in particular when the true MTD is higher than the prior MTD. For example, consider two cases: in case 1, the prior sets dose level 3 as the MTD while the true MTD is dose level 5; and in case 2, the prior sets dose level 1 as the MTD while the true MTD is dose level 3. Although both priors are misspecified, iBOIN and iKeyboard designs have lower probability

to identify the true MTD in case 1 than case 2. This is because in case 2, the prior MTD is dose level 1, when data cumulates, they eventually will override the prior and escalate to the MTD. In contrast, in case 1, because the MTD is dose level 5, the sample size is often exhausted before enough data are cumulated at dose level 3 (i.e., prior MTD) to override the prior.

This observation motivated us to propose a robust prior, which is useful when there is large uncertainty on the prior information. Given the elicited skeleton (q_1, \dots, q_J) with dose level j^* as the prior estimate of the MTD (i.e., $q_{j^*} = \phi$), the robust prior is the same as the prior described above if when $j^* < J/2$, but modify the PESS to $(n_{01}, \dots, n_{0j^*}, 0, \dots, 0)$ when $j^* \geq J/2$. In other words, when prior MTD $j^* \geq J/2$, the robust prior uses informative prior information for the dose up to the prior estimate of the MTD, and after that uses non-informative prior. This modification facilitates overriding prior when the data conflict with the prior, and thus alleviates the impact of the prior misspecification. Our simulation study described later shows that iBOIN and iKeyboard designs are robust to moderate misspecification of priors, and using robust prior proves their robustness when the prior is severely misspecified.

3.2.5 Choose PESS

PESS should be chosen to reflect the appropriate amount of prior information to be incorporated, which depends on the reliability of prior information and varies from trial to trial. When there is strong evidence that the prior is most likely correctly specified, it is appropriate to use a large PESS to borrow more information; when there is large uncertainty on whether the prior is most likely correctly specified, we may use a small PESS to avoid bias. In practice, there is often sizable uncertainty on the reliability of the prior information. Thus, PESS should be chosen carefully to achieve an appropriate balance between design performance and robustness. Using a large PESS improves the design performance (i.e., the accuracy to identify the MTD) when the prior is correctly specified, but may lead to substantial loss of performance when the prior is misspecified. Based on numerical study, we recommend $\text{PESS} \in [1/3(N/J), 1/2(N/J)]$ as default value which improves trial performance while maintaining reasonably robust. For example, when $J = 5$ and $N = 30$, the recommended value for PESS is $n_{0j} = 2$ or 3 (i.e., across 5 doses, the total PESS is 10 or 15). The value of n_{0j} can be further calibrated by simulation using the software described below.

3.3 Software

We have developed online software “BOIN Suite”, freely available at <https://www.trialdesign.org>, to allow users to design trials, conduct simulation, and generate protocol template. The software has an intuitive graphical user interface and rich documents to help with navigating through the process, see Figure A3.1 for the user interface of the software. A trial can be easily designed by the following 3 steps:

Step 1. Specify the design parameters, e.g., sample size, cohort size, target DLT probability, skeleton, PESS, etc.

Step 2. Use the software to produce decision table and design diagram, and conduct simulation to obtain operating characteristics of the design. The software also generates sample texts and protocol template to facilitate the protocol write-up.

Step 3. Use the design decision table to conduct the trial.

After a trial is completed, the app can be used to select the MTD.

3.4 Simulation

We conducted extensive simulation to evaluate the operating characteristics of the proposed designs. We assume $J = 5$ doses and the target DLT probability $\phi = 0.3$. The maximum sample size is $N = 30$ with a cohort size of 3. We considered the CRM with an informative prior (denoted as iCRM), iBOIN and iKeyboard, as well as their counterparts with non-informative prior. We also considered BOIN and Keyboard designs with the robust prior, denoted as iBOIN_R and iKeyboard_R, respectively. For iBOIN and iKeyboard, we set PESS $n_{0j} = 3$ for $j = 1, \dots, 5$, and for iCRM, the prior is chosen such that the PESS at the prior MTD is 3. All the designs use the same skeletons (i.e., the prior DLT probabilities), provided in Table 3.2.

3.4.1 Fixed scenarios

We evaluated the performance of the designs in ten scenarios, as shown in Table 3.2. In the first five scenarios, the MTD is located at dose level 1, 2, 3, 4 and 5, respectively; and the prior MTD is correctly specified and matches the true MTD. To reflect the practice, we do not assume that the prior (at each dose level) exactly matches the truth. Here, we call a prior correctly specified if the prior MTD matches the true MTD. Scenarios 6-10 consider the cases that the prior

is mis-specified. Specifically, in scenarios 6 and 7, the prior MTD is one level off the true MTD, and in scenarios 8-10, the prior MTD is two levels off the true MTD.

Table 3.2: Ten dose-toxicity scenarios with target DLT probability $\phi = 0.30$. The prior MTDs are correctly specified in scenarios 1-5 and misspecified in scenarios 6-10.

	Dose level					Dose level				
	1	2	3	4	5	1	2	3	4	5
	<u>Scenario 1</u>					<u>Scenario 6</u>				
True Pr(DLT)	0.30	0.42	0.50	0.60	0.65	0.09	0.12	0.15	0.30	0.45
Prior Pr(DLT)	0.30	0.42	0.54	0.64	0.73	0.01	0.04	0.10	0.19	0.30
	<u>Scenario 2</u>					<u>Scenario 7</u>				
True Pr(DLT)	0.15	0.27	0.40	0.50	0.65	0.08	0.15	0.31	0.45	0.55
Prior Pr(DLT)	0.19	0.30	0.42	0.54	0.64	0.19	0.30	0.42	0.54	0.64
	<u>Scenario 3</u>					<u>Scenario 8</u>				
True Pr(DLT)	0.08	0.15	0.31	0.45	0.55	0.08	0.15	0.31	0.45	0.55
Prior Pr(DLT)	0.10	0.19	0.30	0.42	0.54	0.01	0.04	0.10	0.19	0.30
	<u>Scenario 4</u>					<u>Scenario 9</u>				
True Pr(DLT)	0.09	0.12	0.15	0.30	0.45	0.04	0.08	0.10	0.18	0.27
Prior Pr(DLT)	0.04	0.10	0.19	0.30	0.42	0.04	0.09	0.30	0.40	0.45
	<u>Scenario 5</u>					<u>Scenario 10</u>				
True Pr(DLT)	0.05	0.08	0.10	0.14	0.30	0.08	0.10	0.28	0.40	0.45
Prior Pr(DLT)	0.01	0.04	0.10	0.19	0.30	0.30	0.42	0.54	0.64	0.73

Table 3.3 shows the results, including (1) percentage of correct selection (PCS), defined as the percentage of simulated trials in which the MTD is correctly identified; (2) percentage of patients treated at MTD; (3) percentage of patients treated above the MTD; (4) risk of overdosing, defined as the percentage of simulated trials that assigned 50% or more patients to the doses above the MTD; and (5) risk of poor allocation: defined as the percentage of simulated trials that assigned less than 6 patients to the MTD. As noted by Zhou et al. [125], metrics (4) to (5) measure the reliability of the design, i.e., the likelihood of a design demonstrating extreme problematic behaviors (e.g., treating 50% or more patients at toxic doses, or fewer than 6 patients at the MTD), which are of great practical importance. Note that the percentage of patients overdosed (i.e., metric (3)) does not cover the risk of overdosing (i.e., metric (4)). Two designs can have a similar percentage of patients overdosed, but rather different risks of overdosing 50% of the patients.

Table 3.3: Operating characteristics of iCRM, iBOIN and iKeyboard, in comparison with their counterparts with non-informative priors. iBOIN_R and iKeyboard_R are iBOIN and iKeyboard using robust prior.

Design	% Correct selection	% Patients at MTD	% Patients above MTD	Risk of overdosing	Risk of poor allocation
Scenario 1					
CRM	54.8	59.9	27.8	23.2	12.2
iCRM	63.1	65.2	24.9	19.4	9.8
BOIN	59.2	59.6	29.0	23.6	10.2
iBOIN	64.2	66.2	22.4	12.8	4.5
iBOIN _R	64.2	66.2	22.4	12.8	4.5
Keyboard	59.2	59.3	29.3	23.6	10.2
iKeyboard	64.2	50.7	39.6	34.2	17.8
iKeyboard _R	64.2	50.7	39.6	34.2	17.8
Scenario 2					
CRM	51.6	36.1	7.7	29.5	25.2
iCRM	53.3	42.4	5.6	23.7	16.8
BOIN	50.6	41.1	6.0	23.0	17.1
iBOIN	57.8	47.6	3.7	10.4	8.6
iBOIN _R	57.8	47.6	3.7	10.4	8.6
Keyboard	50.2	41.1	6.0	23.0	16.7
iKeyboard	59.6	37.8	6.6	35.1	23.6
iKeyboard _R	59.6	37.8	6.6	35.1	23.6
Scenario 3					
CRM	57.2	37.8	22.1	17.3	21.6
iCRM	60.2	40.4	20.9	15.3	18.8
BOIN	52.3	35.7	17.0	7.9	19.2
iBOIN	59.8	41.3	14.5	3.5	10.9
iBOIN _R	58.9	38.2	17.7	9.1	15.2
Keyboard	52.4	35.7	17.1	7.9	18.9
iKeyboard	62.5	35.8	28.8	18.7	19.1
iKeyboard _R	59.7	35.6	29.0	19.4	19.7
Scenario 4					
CRM	52.0	30.0	15.3	10.3	33.4
iCRM	56.6	33.8	14.4	8.6	26.5
BOIN	51.5	28.6	13.1	1.2	24.6
iBOIN	59.7	36.1	12.1	0.6	12.8
iBOIN _R	57.6	32.4	15.7	3.2	19.0
Keyboard	52.1	28.7	13.1	1.2	24.6
iKeyboard	65.1	31.7	25.6	11.2	18.9

Table 3.3 Continued:

Design	% Correct selection	% Patients at MTD	% Patients above MTD	Overdose (%)	% Poor allocation
iKeyboard _R	62.4	31.7	25.6	11.2	18.9
Scenario 5					
CRM	72.7	38.6	0	0	23.4
iCRM	75.8	41.7	0	0	19.7
BOIN	71.0	35.2	0	0	16.8
iBOIN	76.8	42.2	0	0	9.6
iBOIN _R	76.8	42.2	0	0	9.6
Keyboard	71.0	35.2	0	0	16.8
iKeyboard	75.8	47.1	0	0	5.2
iKeyboard _R	75.8	47.1	0	0	5.2
Scenario 6					
CRM	50.7	29.9	14.6	9.3	32.9
iCRM	57.3	33.8	17.0	13.0	28.2
BOIN	51.5	28.6	13.1	1.2	24.6
iBOIN	58.6	35.5	18.4	3.8	11.8
iBOIN _R	58.6	35.5	18.4	3.8	11.8
Keyboard	52.1	28.6	13.1	1.2	24.6
iKeyboard	56.5	35.4	27.9	11.2	17.9
iKeyboard _R	59.5	30.8	27.9	11.2	17.9
Scenario 7					
CRM	58.0	38.1	21.7	17.3	21.8
iCRM	59.8	38.0	18.3	13.4	21.3
BOIN	52.3	35.6	17.0	7.9	19.2
iBOIN	61.6	33.0	10.9	2.2	14.8
iBOIN _R	61.6	33.0	10.9	2.2	14.8
Keyboard	52.4	35.7	17.1	7.9	18.9
iKeyboard	56.4	45.4	15.3	3.6	6.8
iKeyboard _R	56.4	45.4	15.3	3.6	6.8
Scenario 8					
CRM	57.8	37.0	21.6	17.2	22.9
iCRM	58.7	41.5	25.5	21.3	20.3
BOIN	52.3	35.6	17.0	7.9	19.2
iBOIN	54.3	36.2	28.6	19.9	19.8
iBOIN _R	54.3	36.2	28.6	19.9	19.8
Keyboard	52.4	35.7	17.1	7.9	18.9
iKeyboard	46.8	33.0	37.8	34.4	25.2

Table 3.3 Continued:

Design	% Correct selection	% Patients at MTD	% Patients above MTD	Overdose (%)	% Poor allocation
iKeyboard _R	46.8	33.0	37.8	34.4	25.2
Scenario 9					
CRM	67.5	36.5	0	0	30.0
iCRM	64.8	35.3	0	0	31.8
BOIN	69.4	33.8	0	0	22.4
iBOIN	51.4	25.7	0	0	35.3
iBOIN _R	68.8	36.7	0	0	21.2
Keyboard	69.4	33.8	0	0	22.4
iKeyboard	58.7	39.8	0	0	17.8
iKeyboard _R	71.7	39.8	0	0	17.8
Scenario 10					
CRM	54.2	36.5	0.0	28.3	26.1
iCRM	60.6	40.0	0.0	15.2	18.1
BOIN	53.1	37.5	0.1	14.6	17.2
iBOIN	65.5	36.1	0.0	3.1	13.4
iBOIN _R	65.5	36.1	0.0	3.1	13.4
Keyboard	52.6	37.5	0.1	14.6	17.1
iKeyboard	66.5	37.9	0.1	3.8	7.0
iKeyboard _R	66.5	37.9	0.1	3.8	7.0

In scenarios 1 to 5, the prior is correctly specified. iCRM and iBOIN outperform their counterparts using non-informative priors. Specifically, compared to the CRM, iCRM improves the PCS and percentage of patients treated at the MTD by 2-8% and 3-6%, respectively. Compared to the BOIN, iBOIN improves the PCS and percentage of patients treated at the MTD by 5-8% and 5-7%, respectively. iCRM and iBOIN yield comparable PCS and the percentage of patients assigned to the MTD, but iBOIN is more reliable with lower risk of overdoing. For example, in scenarios 2 and 3, the risk of overdosing for iBOIN is about half of that of iCRM. Compared to its non-informative counterpart, iKeyboard has 5-10% increase in PCS, but the percentage of patients treated at the MTD is often lower and the risk of overdosing is substantially increased by more than 10% in most scenarios. The proposed robust prior works well. iBOIN_R and iKeyboard_R yield the performance very similar to iBOIN and iKeyboard, respectively.

Scenarios 6 and 7 consider the cases, where the prior is misspecified with the prior MTD one level off the true MTD. iCRM and iBOIN are robust to this moderate prior misspecification and

outperform their counterparts. For example, in scenario 6, the PCS of iCRM and iBOIN are 57.3% and 58.6%, respectively, about 6.6% and 7.1% higher than CRM and BOIN. Compared to iCRM, iBOIN has lower risk of overdosing. Scenarios 8 and 9 examine the cases where the prior is severely misspecified with the prior MTD two levels off the true MTD. When the prior MTD is higher than the true MTD (i.e., scenario 8), iCRM and iBOIN perform well, yielding performance comparable to their non-informative counterparts. The PCS of the iKeyboard is lower than keyboard. When the prior MTD is lower than the true MTD (i.e., scenario 9), the prior misspecification has more impact on the performance of the designs. The PCS of iCRM and iBOIN is lower than their non-informative counterparts. Scenario 9 is a difficult scenario because the true MTD is the highest dose. The sample size often exhausted before enough data are accumulated to overcome the misspecified prior to reach the highest dose (i.e., MTD). In this scenario, the iCRM performs better than iBOIN because the iCRM tends to escalate dose more aggressively, as demonstrated by its relatively high risk of overdosing. The proposed robust prior addressed this issue. iBOIN_R yield higher PCS that is comparable to the BOIN using non-informative prior. In the case that the prior MTD is two levels lower than the true MTD but the true MTD is not the highest dose, the iCRM and iBOIN outperform their non-informative counterparts (see scenario 10).

3.4.2 Random scenarios

To validate the above results, we repeated the simulation using a large number of scenarios randomly generated using a pseudo-uniform algorithm [16]. Two large sets of random scenarios were constructed. The first set was used to examine the operating characteristics of the designs when the prior is correctly specified. We generated 2000 random scenarios with MTD located at dose level 1, 2, 3, 4, and 5, with equal probability, and assume that the prior MTD is correctly specified for each of the scenarios. The second set was used to evaluate the performance of the designs when prior is misspecified. We considered two types of misspecification: the prior MTD is one level off the true MTD, and the prior MTD is two levels off the true MTD. For each type of misspecification, we generated 4000 random scenarios with half of them having the prior MTD (one or two levels) lower than the true MTD, and the other half having the prior MTD (one or two levels) higher than the true MTD. We simulated 2000 trials for each scenario. Details on random scenarios generation and configuration are provided in section A3.0.3 of the Appendix.

Figure 3.2 shows the simulation results when the prior MTD is correctly specified, which are generally consistent with the results based on the fixed scenarios. That is, iCRM and iBOIN

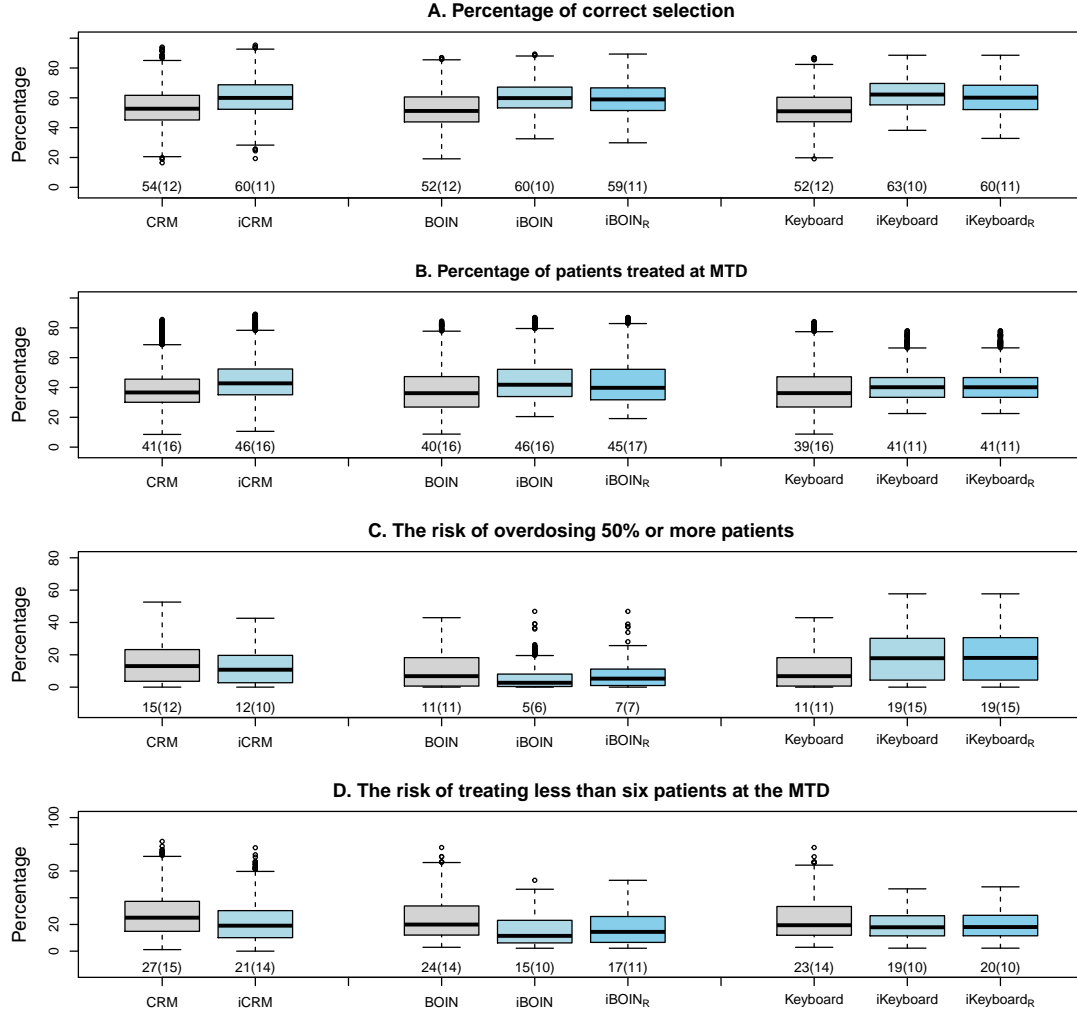


Figure 3.2: Operating characteristics of iCRM, iBOIN, and iKeyboard, in comparison to their counterparts with non-informative priors, under 2000 random scenarios when the prior is *correctly specified*. iBOIN_R and iKeyboard_R are iBOIN and iKeyboard using robust prior, respectively. The number under each boxplot is the average value with the standard deviation shown in parenthesis.

outperform their non-informative counterparts with higher PCS and higher percentage of patients assigned to the MTD. For example, averaging over 2000 random scenarios, the PCS of iCRM is 6% higher than that of the CRM, and the PCS of iBOIN is 8% higher than that of BOIN. iCRM and iBOIN yield similar PCS and the percentage of patients to the MTD, but iBOIN has lower risk of overdosing and poor allocation. iKeyboard design yields higher PCS than its non-informative counterpart, but increases the risk of overdosing due to its aggressive dose escalation.

Figure 3.3 shows the simulation results when prior is misspecified by one dose level. iCRM and iBOIN are robust to such moderate prior misspecification. The PCS of iCRM and iBOIN are both 56%, offering 3% and 5% improvement over their non-informative counterparts, respectively.

The risk of overdosing and poor allocation of iCRM are respectively 5% and 6% higher than the iBOIN. iKeyboard design offered 3% improvement over its non-informative counterpart, but the risk of overdosing is 8% higher. When the prior is severely misspecified with prior MTD being two levels off the true MTD (see Figure 3.4), iCRM is more robust than the iBOIN, however, by using the proposed robust prior, iBOIN_R shows competitive performance. In addition, when the prior is likely to be severely misspecified, at the first place we should avoid using prior information and non-informative prior is a more sensible choice.

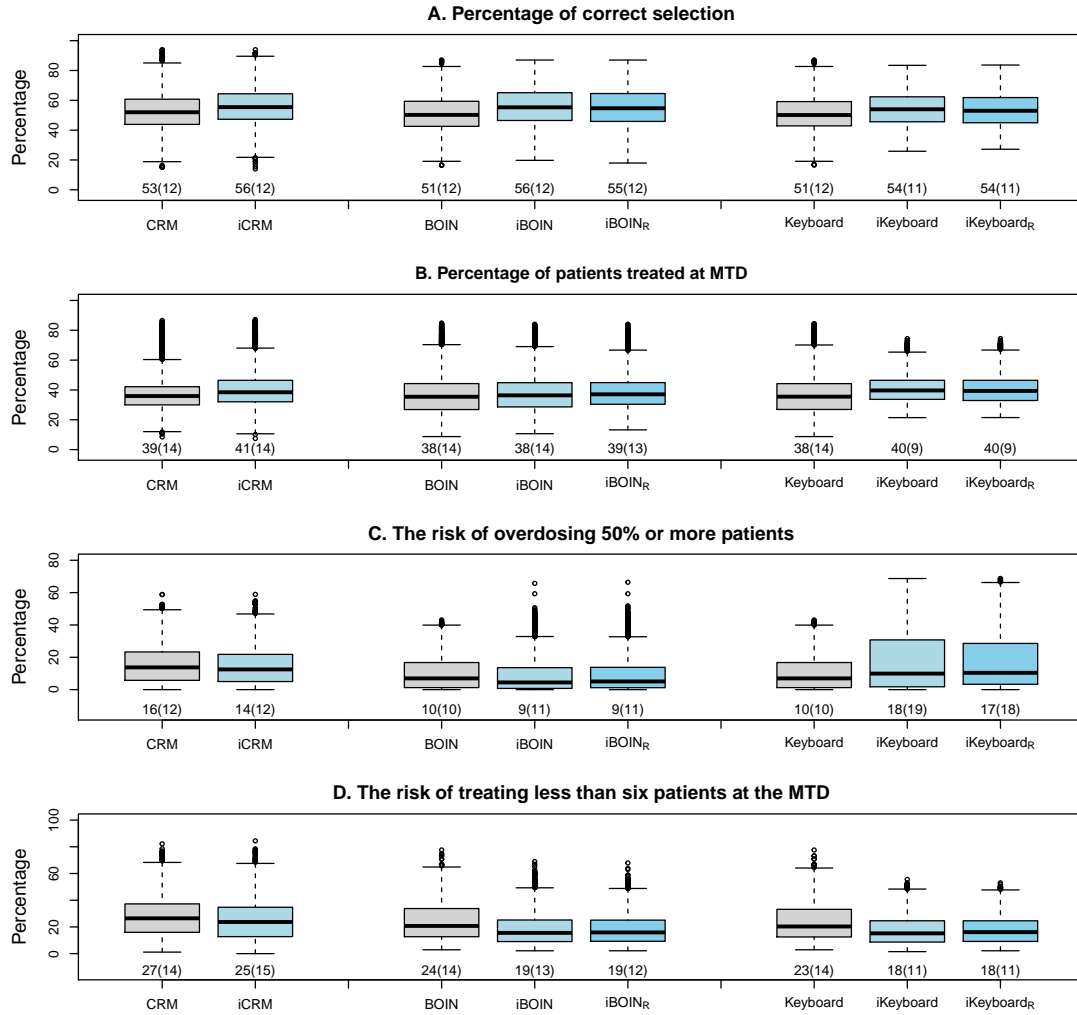


Figure 3.3: Operating characteristics of iCRM, iBOIN and iKeyboard, in comparison to their counterparts with non-informative priors, under 4000 random scenarios when the prior MTD is *one dose off* the true MTD. iBOIN_R and iKeyboard_R are iBOIN and iKeyboard using robust prior, respectively. The number under each boxplot is the average value with the standard deviation shown in parenthesis.

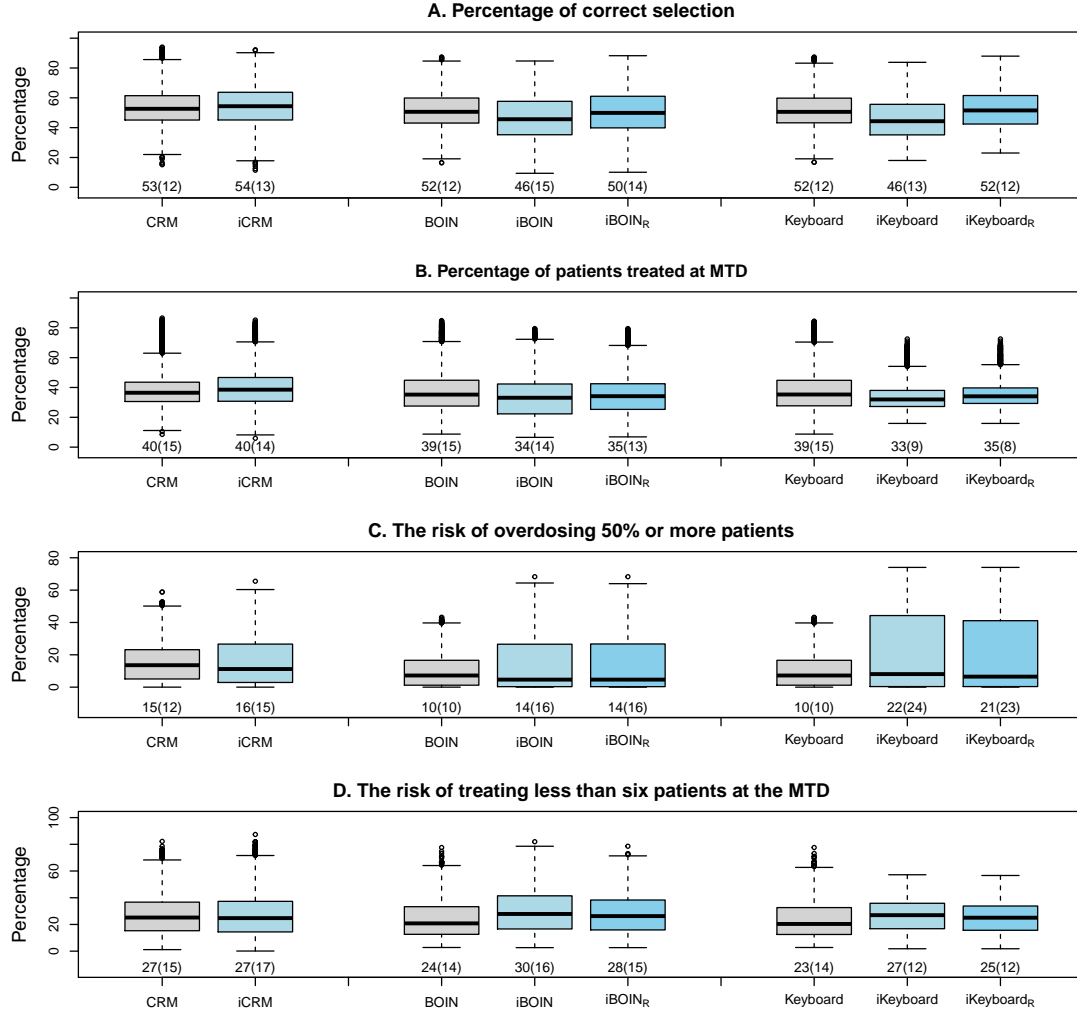
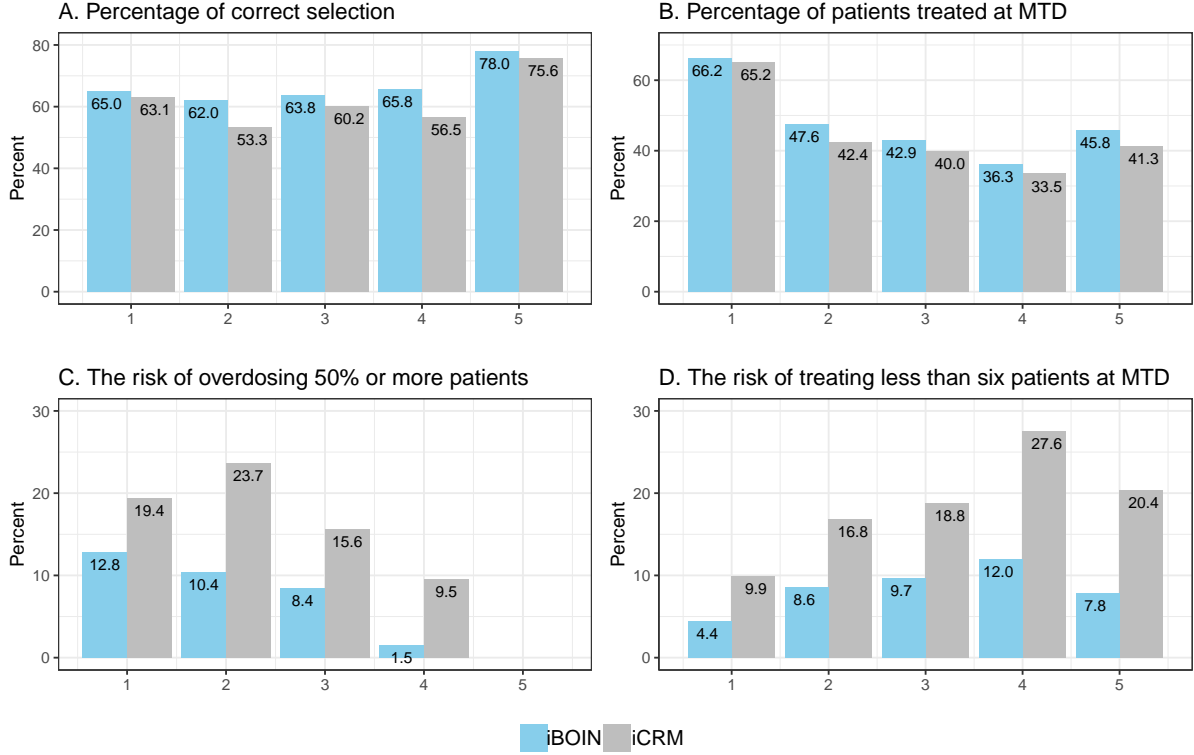


Figure 3.4: Operating characteristics of iCRM, iBOIN and iKeyboard, in comparison to their counterparts with non-informative priors, under 4000 random scenarios when the prior MTD is *two doses off* the true MTD. iBOIN_R and iKeyboard_R are iBOIN and iKeyboard using robust prior, respectively. The number under each boxplot is the average value with the standard deviation shown in parenthesis.

3.4.3 Unequal prior information across doses

Lastly, we briefly investigated the case that different amount of prior information is available for different doses. We assume that more prior data are available at lower dose than higher doses, and more prior data are available around the prior MTD, as we often observed in (historical) phase I trials. As described in section 3.2, iBOIN can easily accommodate that by specifying different PESS at different doses. Figure 3.5 shows the simulation results under scenarios 1 to 5. We control the total PESS over 5 doses are the same for iBOIN and CRM. We see that, compared to the CRM, iBOIN offers higher PCS and allocates higher percentage of patients at the MTD, as well as lower



PESS for the first five scenarios

Scenario	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5
1	6	3	3	2	1
2	3	7	3	1	1
3	2	4	7	1	1
4	1	2	4	7	1
5	2	2	2	4	5

Figure 3.5: Operating characteristics of iBOIN and iCRM when different amount of prior information (i.e., PESS) is available for different doses under scenarios 1 to 5.

risk of overdosing and poor allocation. The CRM does not allow specifying dose-specific PESS as it uses a single parameter to control prior information in all doses, thus cannot take full advantage of the prior information.

3.5 Summary

In this chapter, we propose a unified framework to incorporate historical data or real-world evidence to improve the efficiency of phase I trial designs, especially model-assisted designs. By using skeleton and PESS, our method is intuitive and easy to interpret. More importantly, our approach maintains the hallmark of the model-assisted design: simplicity—the dose escalation/de-escalation rule can be tabulated prior to the trial conduct. For example, implementing the proposed

iBOIN only involves simple comparison of the number of DLT observed at the current dose with the prespecified dose escalation and de-escalation boundaries (e.g., Table 3.1). Extensive simulation studies show that the proposed method, in particular iBOIN, can effectively incorporate prior information and yield comparable performance as the model-based CRM design and greater reliability. Moreover, iBOIN is more transparent and easier to implement. In addition, iBOIN is more flexible and allows specifying dose-specific prior information to more accurately reflect available prior information. The iBOIN design is generally robust to prior misspecification. When there is high likelihood that prior is severely misspecified, the proposed robust prior can be used with iBOIN to enhance its robustness. Actually, in this case, there is little rationale to incorporate prior information and it may be more appropriate to use non-informative prior. When non-informative prior is used, iBOIN becomes the standard BOIN. Freely available software is provided at www.trialdesign.org to facilitate the use of proposed designs.

A Appendix

A3.0.1 Determining informative prior for BOIN

Suppose at dose level j , the prior estimate of DLT probability is q_j with PESS of n_0 . This prior information can be transformed into the prior distribution of the three hypothesis employed by BOIN (i.e., $H_{1j} : p_j = \phi$, $H_{2j} : p_j = \phi_1$, $H_{3j} : p_j = \phi_2$) as follows: for $k = 1, 2$ and 3 ,

$$\begin{aligned}
 \pi_{kj} &= \Pr(H_{kj} \mid n_0, q_j) \\
 &= \sum_{x=0}^{n_0} \Pr(H_{kj} \mid x) \Pr(x \mid n_0, q_j) \\
 &= \sum_{x=0}^{n_0} \frac{\Pr(x \mid H_{kj}) \Pr(H_{kj})}{\sum_{k'=1}^3 \Pr(x \mid H_{kj}) \Pr(H_{kj})} \Pr(x \mid n_0, q_j) \\
 &= \sum_{x=0}^{n_0} \frac{\Pr(x \mid H_{kj})}{\sum_{k'=1}^3 \Pr(x \mid H_{kj})} \Pr(x \mid n_0, q_j). \\
 &= \sum_{x=0}^{n_0} \frac{\phi_k^x (1 - \phi_k)^{n_0-x}}{\sum_{k'=1}^3 \phi_{k'}^x (1 - \phi_{k'})^{n_0-x}} \binom{n_0}{x} q_j^x (1 - q_j)^{n_0-x}.
 \end{aligned}
 \tag{A3.1}$$

By doing so, the prior information is incorporated into the dose escalation and de-escalation boundaries, as given by equation (2.3).

A3.0.2 iBOIN Shiny app interface

Trial Setting

Simulation

Trial Protocol

Select MTD

User guide for the iBOIN App

Doses

Number of doses:

Starting dose level:

5

1

Target Probability

Target Toxicity Probability ϕ :

0.3

☒ Use the default alternatives to minimize decision error (recommended).

Sample Size

Cohort size:

Number of cohort:

3

10

Stop trial if number of patients assigned to single dose reaches:

12

Perform accelerated titration:

☒ No ☐ Yes

Overdose Control

Eliminate dose j if $Pr(p_j > \phi \mid data) > p_E$

Use the default cutoff (recommended) $p_E =$

0.95

☐ Check the box to impose a more stringent safety stopping rule:

Prior Specification

Enter prior toxicity probability and effective sample size (ESS) at each dose level:

	D1	D2	D3	D4	D5
Probability	0.10	0.19	0.30	0.42	0.54
ESS	2.00	2.00	2.00	2.00	2.00

Get Flow Chart and Decision Table

Figure A3.1: User interface of iBOIN software

A3.0.3 Random scenario configuration

A3.0.4 Generate random scenarios where prior MTD is correctly specified

To examine how the informative designs perform, we generated 2000 random scenarios with MTD located at dose level 1, 2, 3, 4, and 5, with equal probability. The random scenarios were generated using the following pseudo-uniform algorithm [16]. Given a target DLT probability ϕ and J dose levels,

1. Select one of the $j \in (1, \dots, J)$ with probability $1/J$.
2. Sample $M \sim \text{Beta}(\max\{J - j, 0.5\}, 1)$.
3. Repeatedly sample J toxicity probabilities uniformly on $[0, B]$ until these correspond to a scenario in which dose level j is the MTD, where $B = \phi + (1 - \phi) \times M$ is the upper bound of DLT probability.

In these scenarios, the MTD is the dose with the DLT probability closest to the target ϕ but not necessarily equal to the target ϕ . It is possible to obtain scenarios in which all the doses have DLT probabilities below or above the target ϕ . To ensure that MTD is uniquely and meaningfully defined, we required that the true DLT probability of the MTD is within $[\phi - 0.05, \phi + 0.05]$, and the distance between the MTD and its adjacent doses is greater than 0.05 and less than 0.3, i.e., $0.05 < p_{j+1} - p_j < 0.3$ and $0.05 < p_j - p_{j-1} < 0.3$. The generating process stops until we obtained 2000 random scenarios that satisfied the specification. Figure A3.2 shows 50 scenarios from the 2000 random scenarios generated. The plot shows that the scenarios cover a wide range of possible dose-toxicity scenarios that we may encounter in practice. Each of 2000 the scenarios has their prior MTD correctly specified. The five prior skeletons utilized are presented in scenarios 1-5 in Table 3.2.

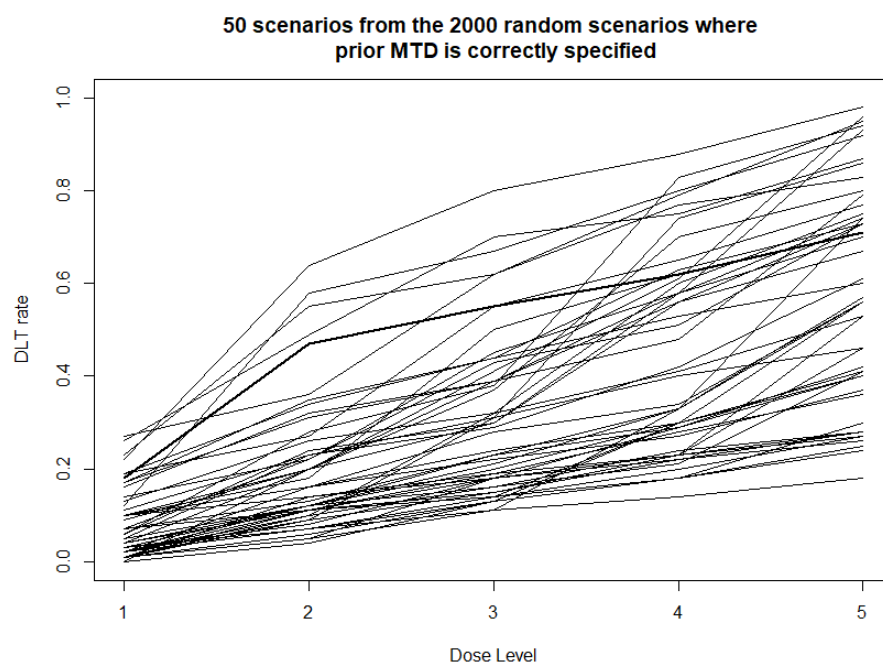


Figure A3.2: 50 randomly selected scenarios from the 2000 scenarios generated

A3.0.5 Generate random scenarios with different levels of mis-specification

To assess the performance of the informative designs when prior is mis-specified (i.e., prior MTD is not corresponding to the true MTD), we conduct extensive simulation for random scenarios with different levels of severity for mis-specification. Below is the configuration of the scenarios.

1. *The prior MTD is one dose below the true MTD.* Generate 2000 scenarios with true MTD located at dose level 2, 3, 4, and 5 with equal probability. The corresponding prior skeletons are in scenarios 1, 2, 3, and 4 in Table 3.2.
2. *The prior MTD is one dose above the MTD.* Generate 2000 scenarios with true MTD located at dose level 1, 2, 3, and 4 with equal probability. The corresponding prior skeletons are in scenarios 2, 3, 4, and 5 in Table 3.2.
3. *The prior MTD is two doses below the true MTD.* Generate 2000 scenarios with true MTD located at dose level 3, 4, and 5 with equal probability. The corresponding prior skeletons are in scenarios 1, 2, and 3 in Table 3.2.
4. *The prior MTD is two doses above the MTD.* Generate 2000 scenarios with true MTD located at dose level 1, 2, and 3 with equal probability. The corresponding prior skeletons are in scenarios 3, 4, and 5 in Table 3.2.

CHAPTER 4

A utility-based Bayesian optimal interval (U-BOIN) Phase I/II design to identify the optimal biological dose for targeted and immune therapies

4.1 Introduction

This chapter is based on an article published in *Statistics in Medicine* in 2019 and entitled “A utility-based Bayesian optimal interval (U-BOIN) phase I/II design to identify the optimal biological dose for targeted and immune therapies” [126], for which I was the first-author. It was a joint work with J. Jack Lee and Ying Yuan. The journal has granted the permission to use this article in the dissertation.

Immunotherapy and targeted therapies have revolutionized cancer treatment. Unlike conventional chemotherapy, immunotherapy drugs do not target the tumor directly. Instead, they work by reactivating the immune system and, hence, reestablishing its capacity to combat tumors. A major class of immunotherapy drugs are monoclonal antibodies, known as immune checkpoint inhibitors (ICI) (e.g., nivolumab, pembrolizumab, and ipilimumab). Checkpoint proteins are receptors on immune cells that can be activated to block immune response, for example checkpoint proteins on T cells (e.g., PD-1 and CTLA-4). The ICI bind to the checkpoint receptors on T cells and release “brake” such that T cells can kill cancer cells, achieving treatment efficacy. Nivolumab and pembrolizumab are PD-1 inhibitors, and ipilimumab is CTLA-4 inhibitor.

Traditional dose-finding designs developed for chemotherapy aim to find the maximum tolerated dose (MTD). The underlying assumption is that both toxicity and efficacy monotonically increase with the dose and, thus, the MTD presents the most efficacious dose that is safe. This assumption, however, is often questionable for immunotherapy and targeted agents. Although it

is reasonable to assume that toxicity increases with the dose, the same is not necessarily true for efficacy. For example, once the checkpoint binding is saturated, further incrementing the ICI dose does not increase treatment efficacy. For some monoclonal antibodies, it is observed that higher doses actually lead to lower efficacy [83]. In addition, while increasing the dose of immunotherapy agents (or targeted biological agents) may improve efficacy, it still may cause substantial toxicity, due to over-activation of the immune system. When the overall benefit is limited, the value of increasing the dose becomes questionable.

To optimize the treatment benefit of immunotherapy and targeted therapy, therefore, it is important to consider toxicity and efficacy simultaneously and their risk-benefit trade-off during dose finding. The objective of dose finding for targeted therapy and immunotherapy is to identify the optimal biological dose (OBD), defined as the dose that has the highest desirability in terms of the risk-benefit trade-off. Our research is motivated by a Phase I/II trial to identify the OBD of a novel humanized anti-TROP2 monoclonal antibody in patients with an advanced solid tumor. Trophoblast cell-surface antigen 2 (TROP2) is associated with increased tumor growth. It is overexpressed in the majority of human epithelial cancers, including esophageal, breast, and lung cancers [73, 75]. Five doses will be studied in the trial. Toxicity will be graded according to the NCI-CTCAE version 4.03, with a 21-day assessment window using the scale of 5 grades. The dose limiting toxicity (DLT) is defined as toxicity with grade 3 or higher. Tumor response will be evaluated using RECIST version 1.1, scored as complete remission (CR), partial remission (PR), stable disease (SD), and progressive disease (PD). Response is defined as CR/PR. TROP2 expression will be measured on day 8, after dosing, to provide a biomarker for measuring the biological activity of the drug.

Here, we develop a utility-based seamless Phase I/II design to find the OBD. We jointly model toxicity and efficacy using a multinomial-Dirichlet model and employ a utility function to measure dose risk-benefit trade-off. We show that the utility approach is flexible and more general, in that it contains the existing marginal toxicity-efficacy trade-off methods as a special case. The design consists of two seamlessly connected stages. In stage I, the Bayesian optimal interval (BOIN) design [65] is used to quickly explore the dose space and collect preliminary toxicity and efficacy data. In stage II, in light of accumulating efficacy and toxicity from both stages I and II, we continuously update the posterior estimate of the utility for each dose after each cohort, and use this information to direct the dose assignment and selection. We refer to the resulting design as the utility-based BOIN (U-BOIN) design. To accommodate the delayed efficacy observed in some

targeted and immunotherapy trials, we extend the U-BOIN design and use the short-term endpoint (e.g., immune activity or other biological activity of targeted agents) to predict the delayed efficacy outcome to facilitate real-time decision making.

Numerous Phase I/II trial designs have been proposed. Thall and Russell [97] developed a Bayesian Phase I/II design that characterizes patient outcomes using a trinary variable that accounts for both toxicity and efficacy. Braun [6] generalized the continual reassessment method (CRM [77]) to accommodate toxicity and efficacy simultaneously. Thall and Cook [96] presented the EffTox design, based on the trade-offs between toxicity and efficacy. Yin et al. [109] proposed a Phase I/II design that uses the odds ratio of the efficacy and toxicity as a measure of desirability. Yuan and Yin [118] described a Phase I/II design that jointly models toxicity and efficacy as time-to-event outcomes. Jin et al. [44] proposed a Phase I/II design that accommodates late-onset toxicity and efficacy. Liu and Johnson [62] proposed a robust Bayesian Phase I/II design, based on a flexible Bayesian dynamic model. Guo and Yuan [31] proposed a personalized Bayesian Phase I/II design that accounts for patient characteristics and biomarker information. Zang et al. [120] proposed several practical Phase I/II trial designs to find the OBD. Liu et al. [61] proposed a Bayesian Phase I/II trial design for immunotherapy that considers the immune response, toxicity, and efficacy. Yuan et al. [113] provided a comprehensive coverage of Phase I/II designs.

Compared to existing Phase I/II designs, the proposed U-BOIN design has several strengths. First, the U-BOIN is simple and easy to implement. It is a model-assisted design in that it uses a model for efficient decision making, but its decision rule for dose escalation/de-escalation can be tabulated and included in the trial protocol before a trial starts [105, 114, 125]. Once the trial is designed, it can be easily implemented using the predetermined decision table. No complicated computation or model estimation is needed. In contrast, most existing Phase I/II designs are model-based designs, and require complicated model fitting and estimation after treating each cohort. Second, the U-BOIN is robust, because it models toxicity and efficacy at each dose independently, without imposing any parametric structure on dose-toxicity and dose-efficacy curves. In contrast, most existing designs assume that the dose-toxicity and dose-efficacy curves follow certain parametric forms (e.g., logistic model), and thus are susceptible to the influence of model mis-specification.

Third, most existing Phase I/II designs use the trade-off in the marginal toxicity and efficacy probabilities of a dose to measure the desirability of the dose, while the U-BOIN design uses the utility, which is not only easier to elicit from physicians, but also more flexible and general. We

prove that the trade-off based on marginal toxicity and efficacy probabilities is a special case of the utility approach. Although utility has been used in previous designs [35, 71, 72], to the best of our knowledge, this work is the first to formally show that the utility approach contains the trade-off based on marginal toxicity and efficacy probabilities as a special case. This provides a theoretical justification for using the utility approach. Lastly, the U-BOIN is capable of accommodating delayed efficacy, whereas most existing designs assume that the efficacy endpoint is quickly observed.

The remainder of this chapter proceeds as follows. Section 4.2 introduces the statistical model, utility function, dose-finding algorithm, and software to implement the U-BOIN design. Section 4.3 uses simulation to compare the U-BOIN design with an existing Phase I/II design, and also assesses the robustness of the design using sensitivity analysis. Section 4.4 provides a brief summary.

4.2 Method

4.2.1 Efficacy-toxicity model

Consider a Phase I/II trial with J doses under investigation. Let $y_E = 0, \dots, R-1$ denote the categorical efficacy endpoint with R levels, where a higher level represents a more desirable treatment response; and $y_T = 0, \dots, Q-1$ denote the categorical toxicity endpoint with Q levels, where a higher level represents a more severe toxicity. The bivariate outcomes (Y_E, Y_T) can be equivalently represented by a single variable Y with $K = R \times Q$ levels, where each level of Y maps to a distinct value of (Y_E, Y_T) .

As an example, consider the conventional setting that both Y_E and Y_T are binary. Let $Y_E = 1$ denote response, 0 otherwise. Similarly, let $Y_T = 1$ denote DLT, 0 otherwise. Then, Y has $K = 4$ levels, with $Y = 1$, if $(Y_E, Y_T) = (0, 1)$; $Y = 2$, if $(Y_E, Y_T) = (0, 0)$; $Y = 3$, if $(Y_E, Y_T) = (1, 1)$; and $Y = 4$, if $(Y_E, Y_T) = (1, 0)$. The value of Y ascribed to each possible (Y_E, Y_T) is not critical, as long as we keep track of the mapping. Without loss of generality, we assume that $Y = 1$ denotes the least favorable clinical outcomes, and $Y = K$ denotes the most favorable clinical outcomes.

Define $\pi_{jk} = \Pr(Y = k \mid d = j)$, $k = 1, \dots, K$ and $j = 1, \dots, J$, with $\sum_{k=1}^K \pi_{jk} = 1$, where d denotes the dose level. We assume that Y follows a Dirichlet-multinomial model as follows:

$$\begin{aligned} Y \mid d = j &\sim \text{Multinomial}(\pi_{j1}, \dots, \pi_{jK}). \\ (\pi_{j1}, \dots, \pi_{jK}) &\sim \text{Dirichlet}(a_1, \dots, a_K), \end{aligned}$$

where $a_1, \dots, a_K > 0$ are hyper-parameters. We set $\sum_{k=1}^K a_k = 1$, such that the prior is vague and equivalent to a prior sample size of 1.

At an interim decision time, assume that n_j patients have been treated at dose $d = j$, among which n_{jk} patients had outcome $Y = k$, where $n_j = \sum_{k=1}^K n_{jk}$. Given the observed interim data $D_j = (n_{j1}, \dots, n_{jK})$, the posterior distribution of $\pi_j = (\pi_{j1}, \dots, \pi_{jK})$ is

$$\pi_j \mid D_j \sim \text{Dirichlet}(a_1 + n_{j1}, \dots, a_K + n_{jK}).$$

4.2.2 Utility

We measure the desirability of the investigational doses using utility. Let ψ_k denote the utility ascribed to outcome $Y = k$, $k = 1, \dots, K$. The utility ψ_k should be elicited from physicians to reflect the risk-benefit trade-off underlying their medical decisions, which can be done using the following procedure:

- (1) Fix the value of the utility for the least desirable outcome $Y = 1$ as $\psi_1 = 0$, and for the most desirable outcome $Y = K$ as $\psi_K = 100$. For example, for binary Y_E and Y_T , the least desirable outcome is $(Y_E = 0, Y_T = 1)$, i.e., (no response, DLT), and the most desirable outcome is $(Y_E = 1, Y_T = 0)$, i.e., (response, no DLT).
- (2) Ask the clinician to use these two utilities as a reference to score the utility values $\psi_2, \dots, \psi_{K-1}$ for the other $K - 2$ possible outcomes $Y = 2, \dots, K - 1$ to quantify the risk-benefit trade-off under each outcome.

Table 4.1 shows three examples of the utility function. The first two examples consider the scenario where both toxicity and efficacy are binary outcomes. Example 1 has utility values $\{\psi_1 = 0, \psi_2 = 30, \psi_3 = 50, \psi_4 = 100\}$ for the outcomes $\{(Y_E = 0, Y_T = 1), (Y_E = 0, Y_T = 0), (Y_E = 1, Y_T = 1), (Y_E = 1, Y_T = 0)\}$. Compared to example 1, example 2 rewards the response (i.e., $Y_E = 1$) more, in the presence of DLT (i.e., $Y_T = 1$), by assigning a larger value to ψ_3 (65 versus 50). This is appropriate for the trial where toxicity can be well managed and response is highly desirable (e.g., leading to long survival). Example 3 shows the case where Y_T has three levels (i.e., minor, moderate, and severe toxicity), and Y_E has also three levels (i.e., PD, SD, and CR/PR).

In our experience, clinicians quickly understand what the utilities mean and provide values for ψ_k 's, since the values reflect the actual clinical practice. After completing this process, simulation should be performed to verify the operating characteristics of the design. In some cases, the simulation results may motivate slight modification of some of the numerical utility values, although

Table 4.1: Examples of utility

(a) Example 1			
	$Y_T = 1$	$Y_T = 0$	
$Y_E = 0$	$\psi_1 = 0$	$\psi_2 = 30$	
$Y_E = 1$	$\psi_3 = 50$	$\psi_4 = 100$	
(b) Example 2			
	$Y_T = 1$	$Y_T = 0$	
$Y_E = 0$	$\psi_1 = 0$	$\psi_2 = 30$	
$Y_E = 1$	$\psi_3 = 65$	$\psi_4 = 100$	
(b) Example 3			
	$Y_T = \text{severe}$	$Y_T = \text{moderate}$	$Y_T = \text{minor}$
$Y_E = \text{PD}$	$\psi_1 = 0$	$\psi_2 = 15$	$\psi_3 = 30$
$Y_E = \text{SD}$	$\psi_4 = 0$	$\psi_5 = 30$	$\psi_6 = 50$
$Y_E = \text{PR/CR}$	$\psi_7 = 15$	$\psi_8 = 45$	$\psi_9 = 100$

Note. PD: partial disease, SD: stable disease, PR: partial remission, CR: complete remission.

such modification typically has little or no effect on the design's operating characteristics. One possible criticism for using the utility values is that they require subjective input. However, we are inclined to view this as a strength, rather than a weakness. The process of specifying the utility requires clinicians to carefully consider the potential risks and benefits of the treatment that underlie their clinical decision making in a more formal way and incorporate that into the trial. In addition, our simulation study and previous studies [31, 61, 72] show that the design is generally not sensitive to the numerical values of the utility, as long as it reflects a similar trend.

Given the values of ψ_k , the true mean utility for dose j is given by

$$(4.1) \quad U_j = \sum_{k=1}^K \psi_k \pi_{jk},$$

Since the true mean utility U_j depends on π_{jk} , which is unknown, we need to estimate it, based on the observed data. Given the interim data $D = \{D_j\}$, the estimate of mean utility is given by

$$(4.2) \quad \hat{U}_j = \sum_{k=1}^K \psi_k E(\pi_{jk} \mid D).$$

For the conventional setting with binary Y_T and Y_E , another common approach to defining the efficacy-toxicity trade-off is directly based on the marginal efficacy probability $\pi_{E,j} = \Pr(Y_E =$

$1|d = j)$ and marginal DLT probability $\pi_{T,j} = \Pr(Y_T = 1|d = j)$, which can be expressed as

$$(4.3) \quad U_j^M = \pi_{E,j} - w\pi_{T,j},$$

where w is a prespecified weight [113]. This trade-off function says that patients are willing to trade an increase of w in the DLT probability for a unit increase in the efficacy probability. If $w = 0$, we obtain the special case that the dose with the highest efficacy is the most desirable. The following theorem shows that this marginal-probability-based approach is a special case of the utility approach described above. The proof is provided in the Appendix.

Theorem 1. Marginal-probability-based trade-off U_j^M , defined in (4.3), is a special case of the utility method defined in (4.1), in the sense that for a prespecified weight w , we can find (ψ_2, ψ_3) , such that $U_j = \xi U_j^M$, where ξ is a non-zero constant.

In the Supplementary Material available at <https://onlinelibrary.wiley.com>, we provide an example to show how to map U_j with U_j^M by choosing appropriate values of (ψ_2, ψ_3) .

Liu and Johnson [62] proposed another marginal-probability-based trade-off function

$$(4.4) \quad U_j^{M2} = \pi_{E,j} - w_1\pi_{T,j} - w_2\pi_{T,j}I(\pi_{T,j} > \rho),$$

where w_1 and w_2 is a prespecified weight, $I(\cdot)$ is an indicator function, and ρ is a prespecified toxicity threshold deemed of substantial concern. Compared to U_j^M in (4.3), this trade-off function is more flexible and allows to impose a higher penalty (i.e., $w_1 + w_2$) when the true DLT probability $\pi_{T,j}$ exceeds the threshold ρ . U_j^{M2} becomes U_j^M when $w_2 = 0$. It is not clear if trade-off function U_j^{M2} can be expressed equivalently in the form of U_j as (4.1). Nevertheless, the proposed U-BOIN design can be used with different types of utility/trade-off, including U_j^{M2} . We plan to incorporate this functionality in our software provided later.

4.2.3 Optimal biological dose (OBD)

To define the OBD, we first define the admissible dose to safeguard patients from toxic or futile doses. As the objective here is to rule out toxic or futile doses, it is more natural to start with the definition of an inadmissible dose. Let $\bar{\pi}_T$ denote the maximum tolerable DLT probability, and $\underline{\pi}_E$ the lowest acceptable response rate. We define that dose j is inadmissible, if it meets either one

or both of the following two criteria:

$$(4.5) \quad (\text{Toxic}) \quad \Pr(\pi_{T,j} > \bar{\pi}_T \mid D) > C_T,$$

$$(4.6) \quad (\text{Futile}) \quad \Pr(\pi_{E,j} < \bar{\pi}_E \mid D) > C_E,$$

where C_E and C_T are probability cutoffs. In general, $C_T = 0.95$ and $C_E = 0.9$ work well, but should be calibrated using simulation to ensure desirable operating characteristics. This can be done easily using the software provided.

The admissible dose is then defined as the dose for which none of the criteria (4.5) and (4.6) is satisfied. When Y_T has more than 2 categories (e.g., grade 0-2, grade 3, and grade 4-5), for the purpose of defining the admissible dose, Y_T can be temporarily collapsed into DLT/no DLT (i.e., DLT = grade 3 and higher). This dichotomization simplifies the definition of the admissible dose and is often adequate to safeguard patients from overly toxic doses. While it is not necessary, more complicated safety criteria could be entertained to accommodate multiple categories. Similar dichotomization (i.e., response/no response) is also applied to Y_E , when it has more than two categories. Note that although we dichotomize Y_T and Y_E for defining the admissible dose, our model and utility are still based on their original scales. We define the OBD as the dose that is admissible and has the highest utility value, i.e.,

$$\text{OBD} = \operatorname{argmax}_{j \in \mathcal{A}}(U_j),$$

where \mathcal{A} denotes the set of admissible doses.

4.2.4 Phase I/II OBD finding algorithm

The U-BOIN design consists of two seamless, connected stages (Figure 4.1). The objective of stage I is to quickly explore the dose space to identify a set of admissible doses that are reasonably efficacious and safe for stage II. In stage I, we conduct dose escalation based on only the toxicity outcome, but efficacy data are also collected and will be used for decision making in stage II. Given the exploratory nature of stage I, if Y_T has more than two categories, we dichotomize it as DLT/no-DLT to facilitate the exploration of the dose space. This is in line with the clinical practice and serves well for the purpose of stage I. Note that for the estimation and finding the OBD (i.e., the primary objective of the trial), we retain the original scale of Y_T .

Stage I dose escalation/de-escalation is guided by the BOIN design [65], which has been

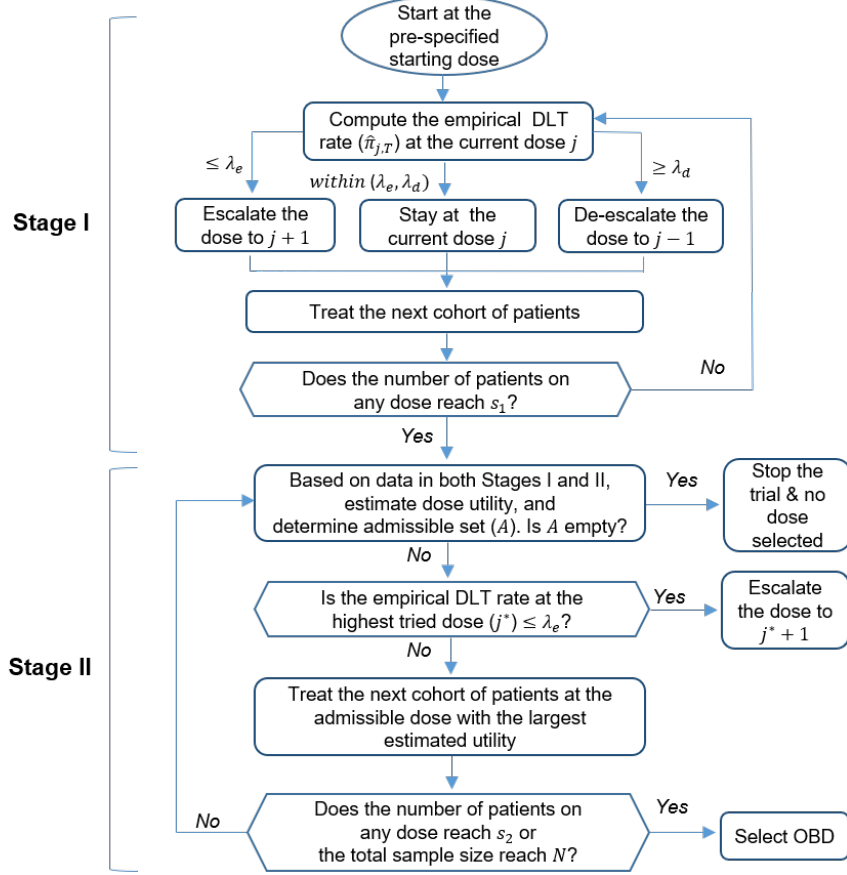


Figure 4.1: Diagram of the U-BOIN design

well-validated and widely used in a variety of oncology trials [1, 51, 54, 104] and treatment agents [41, 52, 66, 82]. Due to very limited data and large uncertainty, for patient safety, we set the target DLT probability $\phi_T = \bar{\pi}_T - 0.05$, slightly lower than the maximum tolerable DLT probability $\bar{\pi}_T$, to ensure that stage I dose exploration concentrates around up to, but not exceeding $\bar{\pi}_T$. Let $\hat{\pi}_{T,j}$ denote the empirical (or maximum likelihood) estimate of $\pi_{T,j}$, given by $\hat{\pi}_{T,j} = m_j/n_j$, where m_j is the number of patients who experienced DLT at the dose level j ; and λ_e and λ_d denote the predetermined optimal escalation boundary and de-escalation boundary. Table 4.2 provides the values of λ_e and λ_d for the commonly used target DLT probability ϕ_T , see Liu and Yuan [65] for the derivation and formula to calculate λ_e and λ_d .

Table 4.2: Dose escalation and de-escalation boundaries of the BOIN design

Boundaries	Target DLT probability (ϕ_T)					
	0.15	0.20	0.25	0.30	0.35	0.40
λ_e (escalation)	0.118	0.157	0.197	0.236	0.276	0.316
λ_d (de-escalation)	0.179	0.238	0.298	0.358	0.419	0.480

The dose-finding algorithm in stage I proceeds as follows.

- A1 Patients in the first cohort are treated at dose level 1 or a prespecified starting dose.
- A2 Suppose j is the current dose, use the following rules to assign a dose to the next cohort of patients.
- Escalate the dose to $j + 1$ if $\hat{\pi}_{T,j} \leq \lambda_e$.
 - De-escalate the dose to $j - 1$ if $\hat{\pi}_{T,j} \geq \lambda_d$.
 - Otherwise, stay at the current dose j .
- A3 Repeat step A2 until the number of patients treated on one of the doses reaches s_1 , and then move to stage II. We recommend $s_1 = 12$ as the default value, while $s_1 = 9$ to 15 generally yields good operating characteristics.

Stage II proceeds as follows:

- B1 Let j^* denote the highest dose level that has been tried. If $\hat{\pi}_{j^*,T} \leq \lambda_e$ and j^* is not the highest dose in the trial, escalate the dose to $j^* + 1$ for treating the next cohort of patients; otherwise, proceed to step B2.
- B2 Given the observed interim data D collected in both stages I and II, determine the admissible dose set \mathcal{A} . If no dose is admissible, terminate the trial and no dose should be selected as the OBD. Otherwise, assign the next cohort of patients to the admissible dose (i.e., $\in \mathcal{A}$) that has the largest posterior mean utility, which can be pre-tabulated.
- B3 Repeat steps B1 and B2 until reaching the prespecified maximum sample size N or the number of patients treated at one of the doses reaches s_2 ($> s_1$), and then select the OBD as the admissible dose (i.e., $\in \mathcal{A}$) that has the largest posterior mean utility. For most trials, a value between 18 to 24 is a reasonable choice for s_2 .

In stage I, following the BOIN design, we impose an overdose control rule as follows: if $\Pr(\pi_{T,j} > \bar{\pi}_T \mid m_j, n_j) > 0.95$ and $n_j \geq 3$, dose level j and higher are eliminated from the trial; the trial is terminated if the lowest dose level is eliminated, where $\Pr(\pi_{T,j} > \bar{\pi}_T \mid m_j, n_j) > 0.95$ is evaluated based on a beta-binomial model with the uniform prior. Once the trial move to stage II, this overdose control rule is seamlessly merged as the inadmissible rule (4.5). In the overdose control rule, we use $\bar{\pi}_T$, rather than ϕ_T , as the DLT probability threshold to ensure that the overdose control rule seamlessly connects with the inadmissible rule.

For stage II step B1, the reason that we perform dose escalation when $\hat{\pi}_{T,j^*} \leq \lambda_e$ is to allow the trial to continue exploring the dose space, given that the highest tried dose is safe, to reduce the risk of being stuck at a local suboptimal dose, due to a large variation caused by a small sample size. Besides the pick-the-winner (PW) approach (i.e., deterministically assigning the next cohort of patients to dose $j \in \mathcal{A}$ that has the largest posterior mean utility), other strategies can also be used to assign patients. For example, we can adaptively randomize the next cohort of patients to dose $j \in \mathcal{A}$, with probability ω_j proportional to its posterior mean utility, i.e.,

$$(4.7) \quad \omega_j = \frac{U_j}{\sum_{j \in \mathcal{A}} U_j}.$$

The adaptive randomization (AR) approach can reduce the risk of being stuck at a suboptimal dose, but as a trade-off, it tends to treat fewer patients at the OBD. Another approach is equal randomization (ER), where the next cohort of patients are assigned to the admissible doses with equal probability, i.e.,

$$(4.8) \quad \omega_j = \frac{1}{\sum_{j \in \mathcal{A}} 1}.$$

We compare the performance of PW, AR, and ER in our simulation study. None of the methods dominates the others. Thus, we generally recommend the PW approach because of its simplicity.

Unlike most existing Phase I/II designs, which require complicated model fitting and estimation after each cohort to make the decision of dose assignment, one prominent advantage of the U-BOIN is that its dose assignment rules can be pre-tabulated in decision tables and included in the trial protocol before the trial starts. To conduct the trial, no complicated calculation is needed. The investigator can simply use the decision tables to determine dose assignment, e.g., determine whether escalation/de-escalation is needed (steps A2 and B1) or identify admissible doses (step B2). Because of this feature, the U-BOIN can be classified as a model-assisted design [105, 114, 125]. Section 1 in the online Supplementary Material (at <https://onlinelibrary.wiley.com>) provides an example of using the decision tables in a hypothetical Phase I/II trial.

4.2.5 Delayed efficacy response

In some trials, efficacy endpoint Y_E requires a long time to be ascertained. In our motivating example, it takes three months to evaluate Y_E . The long assessment window causes a major logistics

issue for decision making in stage II. For example, given that the accrual rate is three patients per month, and that patients are treated in cohorts of three, on average, six new patients will be accrued while waiting to evaluate the previous three patients' outcomes. This begs the question: how can new patients receive timely treatment, when the previous patients' outcomes are not yet observed?

Statistically, this means that D are not fully observed, as some Y_E 's are unavailable. As a result, the mean utility estimate (4.2) and the inadmissible criterion (4.6) cannot be evaluated for making interim decisions. This issue is known as the late-onset or delayed-outcome issue, which has been studied in literature. Cheung and Chappell [14] proposed a weighting method to handle late-onset toxicity for Phase I clinical trials. Yuan and Yin [118] modeled toxicity and efficacy as time-to-event outcomes that accommodate the unobserved as censored events. Liu et al. [64] and Jin et al. [44] treated unobserved outcomes as missing data, and proposed a Bayesian data augmentation approach to predict unobserved toxicity and/or efficacy outcomes to facilitate decision making. Cai et al. [8] took the multiple imputation approach to handle unobserved efficacy outcomes in Phase II trials.

We follow the approach of Cai et al. [8] and use multiple imputation to handle unobserved Y_E . The innovation here is that we use the measure of biological activity, which is routinely recorded in targeted and immunotherapy trials, as an ancillary variable to help predict (or impute) Y_E . Examples of the measure of biological activity include immune response (e.g., CD8+ T cell count) in immunotherapy trials and gene expression related to the pathway targeted by the treatment agent. The measure of biological activity is often quickly observable after drug administration and correlated with the clinical response. Daud et al. [18] showed that the abundance of CD8+ T cells predict response to anti-PD1 therapy, and advocated using the immune activity to predict the likelihood of achieving a clinical response to the PD-1 pathway inhibitor. For ease of exposition, hereafter we use the immune response, denoted by Y_I , as the example to illustrate our approach, but Y_I can be any reasonable biological activity measure predictive of treatment efficacy. We assume that Y_I is quickly observable, thus its value is always available at the time of decision making.

Given the observed value of Y_I , we predict the unobserved Y_E , based on the following scaled logistic regression model:

$$(4.9) \quad \text{logit} \left(\frac{\pi_{E,j}}{\lambda} \right) = \beta_0 + \beta_1 Y_I,$$

where $\pi_{E,j}$ is the probability of efficacy for dose j , β_0 and β_1 are regression parameters, and $0 < \lambda \leq 1$ is a plateau (or scale) parameter used to reflect that the probability of clinical response

often levels out after the immune activity reaches a certain level. Under this model, the probability of efficacy $\pi_{E,j}$ increases with Y_I , and then plateaus at the value of λ when Y_I is sufficiently large. When $\lambda = 1$, it becomes a standard logistic regression model. We do not include dose level j in the model, because the treatment effect of immunotherapy is mediated by the immune response, and thus it is often reasonable to assume that, conditional on Y_I , $\pi_{E,j}$ is independent of j . If this assumption is not plausible in some situations, one can simply add j as a covariate to the model. In addition, more biomarkers, when available, can be added to model (4.9) as predictors to improve the prediction accuracy. One attractive property of our imputation approach is that, when the imputation model (4.9) is misspecified, its impact on the design diminishes over time and eventually goes away when the trial is completed. This is because, as the trial proceeds, more and more patients' Y_E become observed, and accordingly fewer and fewer percentage of Y_E needs prediction. As a result, our method is generally robust to model misspecification, as shown later in the sensitivity analysis.

In terms of prior specifications in model (4.9), following Gelman et al.,[27] we assume that the model parameters $(\lambda, \beta_0, \beta_1)$ are independent and have their own prior distributions. We specify a uniform distribution for the scale parameter $\lambda \sim \text{unif}(0,1)$. The magnitude of the coefficients (β_0 and β_1) could be very large or small, depending on different trials. This makes it difficult to specify standard prior distributions for the coefficients. To tackle this problem, we first standardize the variable Y_I . In our application, Y_I is continuous; we standardize it to have a mean of 0 and a standard deviation of 0.5. We regulate the prior distributions to make sure that a typical change in a covariate should not lead to a dramatic change in the efficacy probability.

After standardizing data, a change of 2.5 on the logit scale can move the probability of a favorable response to the therapy from 0.2 to 0.75. We assume that the effect of the immune response is unlikely to be more dramatic than that. This typically is true for immunotherapy trials, as the efficacy probability is rarely outside of that range. We assign a normal distribution with a mean of 0 and a standard deviation 1.25 for β_0 : $\beta_0 \sim N(0, 1.25^2)$. The parameter β_1 is supposed to be positive, due to the positive relationship between immune response and tumor response, and thus we assign a gamma distribution with a shape parameter of 1 and a rate parameter of 1.2: $\beta_1 \sim \text{Gamma}(\text{shape} = 1, \text{rate} = 1.2)$. The priors ensure that a change in the covariate Y_I from one standard deviation below the mean to one standard deviation above the mean will lead to an absolute change that is mostly less than 2.5 on the logit scale.

Let $f(\beta_0, \beta_1, \lambda)$ denote the joint prior distribution of $(\beta_0, \beta_1, \lambda)$. The posterior distribution

of $(\beta_0, \beta_1, \lambda)$ is given by

$$(4.10) \quad f(\beta_0, \beta_1, \lambda | D) \propto f(\beta_0, \beta_1, \lambda) \prod_{j=1}^J \prod_{i=1}^{n_j} \left(\frac{\lambda e^{\beta_0 + \beta_1 y_{I,ji}}}{1 + e^{\beta_0 + \beta_1 y_{I,ji}}} \right)^{y_{E,ji}} \left(\frac{1 + (1 - \lambda) e^{\beta_0 + \beta_1 y_{I,ji}}}{1 + e^{\beta_0 + \beta_1 y_{I,ji}}} \right)^{1 - y_{E,ji}},$$

where $y_{I,ji}$ is the immune response for the i th patient treated with dose j and $y_{E,ji}$ is the tumor response for this patient.

At an interim decision time, let $Y_{E,obs}$ and $Y_{E,mis}$ denote the observed and missing parts of Y_E . We impute the value of $Y_{E,mis}$ using multiple imputations as follows.

- 1) Conditional on observed data $(Y_{E,obs}, Y_I)$, sample L draws of $(\lambda, \beta_0, \beta_1)$ from their posterior distribution (4.10) using the adaptive rejection Metropolis sampling [29]. A certain number of burn-in iterations are typically needed before collecting the posterior draws. In our simulation, we set the number of burn-in iterations as $L' = L/2$.
- 2) Thin the posterior draws by taking a sample after every L'/H draw, resulting in a total of H sets of posterior draws of $(\lambda, \beta_0, \beta_1)$, denoted as $(\lambda^{(1)}, \beta_0^{(1)}, \beta_1^{(1)}), \dots, (\lambda^{(H)}, \beta_0^{(H)}, \beta_1^{(H)})$.
- 3) Impute $Y_{E,mis}$ for H times, where the h th imputed value is generated by drawing a random sample from $Bernoulli(q)$, where $h = 1, \dots, H$ and $q = \lambda e^{\beta_0 + \beta_1 Y_I} / (1 + e^{\beta_0 + \beta_1 Y_I})$. We denoted the H sets of imputed values as $Y_{E,imp}^{(1)}, \dots, Y_{E,imp}^{(H)}$.

After filling in $Y_{E,mis}$ with each of the H imputed values, we obtained an H imputed complete dataset, $D^{(1)}, \dots, D^{(H)}$, where $D^{(h)}$ is obtained by filling in $Y_{E,mis}$ with $Y_{E,imp}^{(h)}$, $h = 1, \dots, H$. Then, the estimate of mean utility (4.2) is given by

$$\hat{U}_j = \frac{1}{H} \sum_{h=1}^H \sum_{k=1}^4 \psi_k E(\pi_{jk} | D^{(h)}),$$

and the left side of the inadmissible criterion (4.6) is calculated as

$$\Pr(\pi_{E,j} < \pi_E | D) = \frac{1}{H} \sum_{h=1}^H \Pr(\pi_{E,j} < \pi_E | D^{(h)}).$$

The dose finding follows the same algorithm described in section 4.2.4. Little and Rubin [58] suggested that, for practical use, the number of multiple imputation (H) can be sufficient when $H = 5$ or greater. In our simulation, we perform $H = 20$ imputations.

To further improve the efficiency of the design, Y_I can also be used to refine the admissible rule. The rationale is that when Y_E takes a long time to be observed, the inadmissible criterion

(4.6) is not effective for screening out ineffective doses, because a high percentage of Y_E may not be observed at the time of making interim decisions. Since Y_I is quickly observable, it can be used, supplementary to Y_E , to improve the power of identifying effective doses or safeguard patients from ineffective doses. Specifically, define $\mu_{I,j} = E(Y_I | d = j)$, and let $\underline{\mu}_I$ denote the lowest acceptable mean immune response. We define dose j inadmissible if it meets at least one of the toxicity and futility criteria in (4.5) and (4.6), and

$$(4.11) \quad (\text{Insufficient activity}) \quad \Pr(\mu_{I,j} < \underline{\mu}_I \mid D) > C_I,$$

where C_I is a prespecified probability threshold. In other words, if a dose has little activity to activate immune system, it is deemed unpromising and inadmissible. Again, the admissible dose is defined as the dose that is not inadmissible. The posterior probability $\Pr(\mu_{I,j} < \underline{\mu}_I \mid D)$ can be evaluated based on the Bayesian normal model:

$$Y_I \mid d = j \sim N(\mu_{I,j}, \sigma_j^2)$$

$$f(\mu_{I,j}, \sigma_j^2) \propto \sigma_j^{-2}.$$

4.2.6 Software and trial implementation

To facilitate the use of the U-BOIN design, we develop an easy-to-use web-based application, which is freely available at <http://www.trialdesign.org>. Figure 4.2 shows the graphical user interface of the app. Extensive help files are available by clicking on the yellow question mark at the upper right corner of each tab.

To design a trial, follow the steps:

- Specify the design parameters (e.g., dose information, sample size, cohort size, utility function, admissible criteria, etc.), under the **Trial Setting** tab. After parameters are entered, design diagram and decisions tables will be provided under the tab.
- Generate the operating characteristics of the design by supplying different scenarios (under the **Simulation** tab).
- Download trial protocol under **Trial Protocol** tab. The protocol includes (1) a brief description of the U-BOIN design, (2) decision tables based on input under **Trial Setting**, and (3) operating characteristics for the scenarios entered under **Simulation**.

Trial Setting
Simulation
Trial Protocol
Trial Conduct

Dose & Sample Size

Number of doses:

Starting dose level:

Cohort size :

Number of cohorts :

Stage I is completed if the number of patients treated at any dose reaches s_1 , where $s_1 =$

Stage II is completed if the number of patients treated at any dose reaches s_2 , where $s_2 =$

Utility for Risk-benefit Tradeoff

Number of toxicity level:

Number of efficacy level:

Toxicity

Efficacy	Yes	No
No	0	30
Yes	50	100

Admissible Criteria

Maximum tolerable toxicity rate ($\bar{\pi}_T$):

Minimum acceptable efficacy rate (π_E):

A dose is deemed admissible if none of the following conditions for the true toxicity rate (π_T) and true efficacy rate (π_E) hold.

(Criterion for toxicity)

$\Pr(\pi_T > \bar{\pi}_T \mid data) > C_T$, where $C_T =$

(Criterion for futility)

$\Pr(\pi_E < \pi_E \mid data) > C_E$, where $C_E =$

Get Flow Chart and Decision Table

Figure 4.2: The user interface of the U-BOIN software

To conduct a trial, use the decision tables generated under **Trial Setting** or use the **Trial Conduct** tab. The app will return the dose for treating the next cohort of patients if the data provided indicates that the trial is not completed, and it will select the OBD if the trial is completed.

4.3 Numerical study

4.3.1 Simulation A

Simulation A considers the case that Y_T and Y_E are quickly ascertainable. Following our motivating trial, we consider $J = 5$ doses, and the total sample size $N = 54$ patients with $s_1 = 12$. The minimum acceptable efficacy probability is $\pi_E = 0.2$, and the maximum acceptable DLT probability is $\pi_T = 0.30$. We set $\pi_T^* = 0.25$ by using $\delta = 0.05$. For the inadmissible criteria (i.e., equations (4.5) and (4.6)), probability cutoffs are set as $C_T = 0.95$ and $C_E = 0.9$, based on simulation calibration. The elicited utility is presented in Table 4.1 as example 1.

We consider 8 representative scenarios that differ in the shape of the dose-toxicity and dose-efficacy curves, and the location of the OBD. The scenarios are shown in Figure 4.3. In scenarios 1 and 2, the dose-response curve increases with the dose level, and then plateaus. In scenarios 3 and 4, the dose-response curve increases to an optimal point, and then decreases. In scenario 5, the dose-response curve levels off at the first dose level. In scenarios 6, 7, and 8, all dose-response curves monotonically increase. Scenario 6 has the last dose level being the OBD. Scenario 7 has two OBDs, located at dose level 2 and 3. Scenario 8 has no OBD since the doses are either futile or overly toxic. Under each scenario, we generate (Y_T, Y_E) , based on a Gumbel model:

$$\begin{aligned} \Pr(Y_T = y_T, Y_E = y_E \mid d = j) = & (\pi_{E,j})^{y_E} (1 - \pi_{E,j})^{1-y_E} (\pi_{T,j})^{y_T} (1 - \pi_{T,j})^{1-y_T} \\ & + \pi_{E,j} (1 - \pi_{E,j}) \pi_{T,j} (1 - \pi_{T,j}) (-1)^{y_E + y_T} \left(\frac{e^c - 1}{e^c + 1} \right), \end{aligned}$$

where $y_E, y_T \in \{0, 1\}$, and the association parameter c was set as 0.2. The values of $\pi_{E,j}$ and $\pi_{T,j}$ (i.e., the marginal efficacy probability and DLT probability) for each dose level are provided in Table 4.3. We compare the proposed U-BOIN design with the EffTox design[96]. We consider the two most important metrics for comparison: (1) percentage of correct selection (PCS), which is the probability of correctly identifying the OBD, and (2) patient allocation, which referred to the average number of patients assigned to each dose. In the EffTox design, the toxicity-efficacy trade-off is constructed based on three equally desirable trade-off target probabilities: $(\pi_E = 0.2, \pi_T = 0)$, $(\pi_E = 1, \pi_T = 0.7)$, and $(\pi_E = 0.45, \pi_T = 0.3)$. Under each scenario, we performed 2,000 simulations.

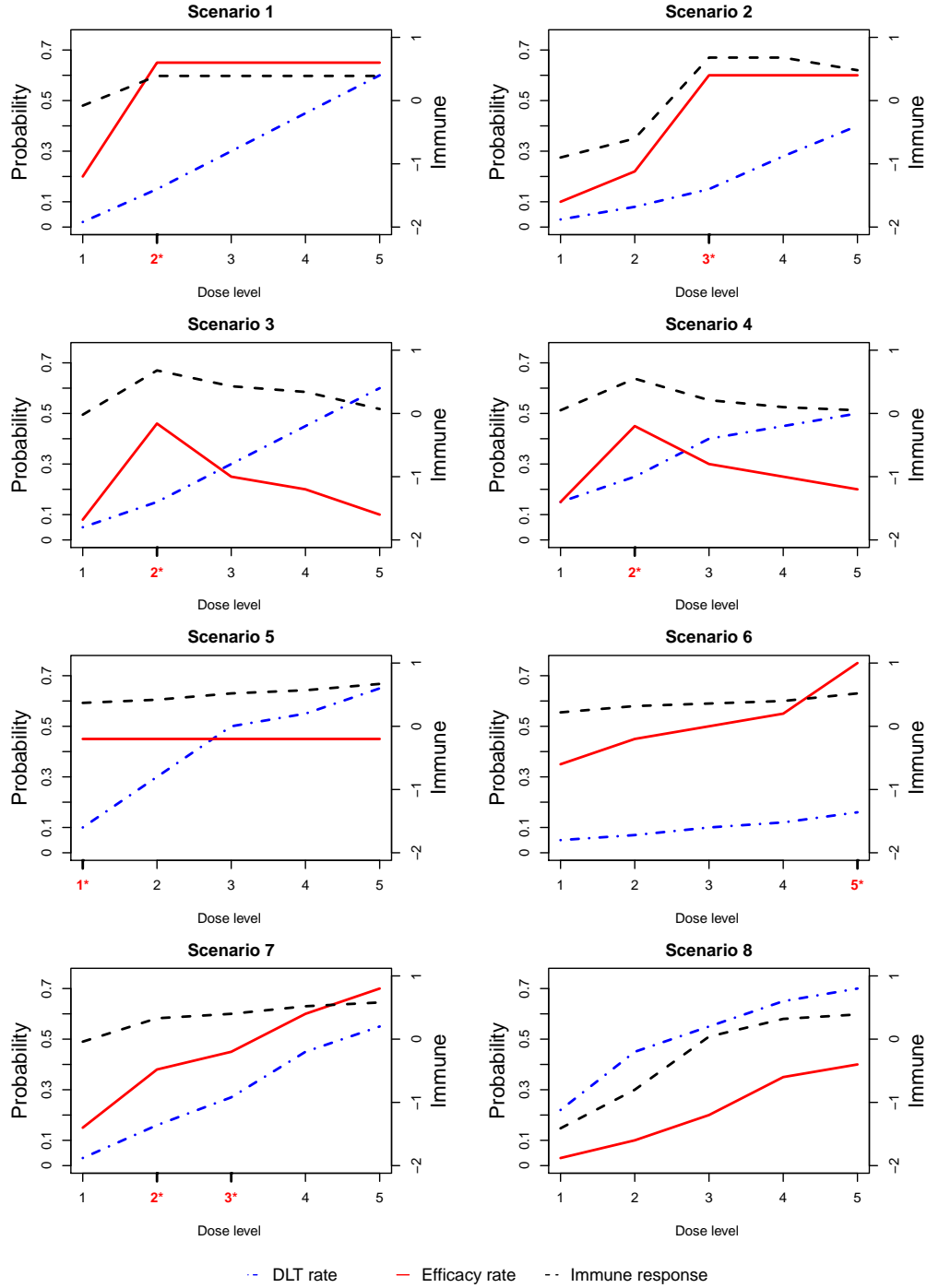


Figure 4.3: Simulation scenarios. The dash-dotted line (blue) is the dose-toxicity curve, the solid line (red) is the dose-efficacy curve, and the dashed line (black) is the dose-immune response curve. The OBD is highlighted by red asterisk in the x-axis. Simulation A considers only the efficacy and toxicity curves, while simulation B considers efficacy, toxicity, and immune response.

Table 4.3: Results of Simulation A, including the selection percentage (Selection %), the average number of patients treated at each dose (No. of patients), and the percentage of early stopping. The optimal biological dose (OBD) is bolded. In scenario 8, the OBD does not exist, and thus the percentage of early stopping is bolded.

Design		Dose Level					% of early
		1	2	3	4	5	stopping
Scenario 1							
	DLT probability	0.02	0.15	0.30	0.45	0.60	
	Efficacy probability	0.20	0.65	0.65	0.65	0.65	
	Utility	43.0	69.0	63.0	56.0	50.0	
EffTox	Selection %	2.0	50.0	45.0	2.0	0.0	0.0
	No. of patients	4.3	22.7	23.8	2.8	0.4	
U-BOIN	Selection %	1.7	72.9	22.4	2.8	0.0	0.2
	No. of patients	6.2	29.9	13.8	3.5	0.5	
Scenario 2							
	DLT probability	0.03	0.08	0.15	0.28	0.40	
	Efficacy probability	0.10	0.22	0.60	0.60	0.60	
	Utility	36.0	43.0	66.0	60.0	55.0	
EffTox	Selection %	0.0	4.0	60.0	29.0	7.0	0.0
	No. of patients	3.4	4.9	26.4	14.2	5.1	
U-BOIN	Selection %	1.1	3.2	65.7	24.9	4.3	0.8
	No. of patients	4.9	7.5	24.4	12.7	4.3	
Scenario 3							
	DLT probability	0.05	0.15	0.30	0.45	0.60	
	Efficacy probability	0.08	0.46	0.25	0.20	0.10	
	Utility	34.0	56.0	37.0	29.0	18.0	
EffTox	Selection %	12.0	70.0	10.0	1.0	0.0	7.0
	No. of patients	10.8	25.5	11.5	2.9	1.0	
U-BOIN	Selection %	1.2	92.2	4.0	0.4	0.0	2.1
	No. of patients	6.6	34.7	8.9	2.6	0.4	
Scenario 4							
	DLT probability	0.15	0.25	0.40	0.45	0.50	
	Efficacy probability	0.15	0.45	0.30	0.25	0.20	
	Utility	36.0	52.0	36.0	32.0	27.0	
EffTox	Selection %	36.0	47.0	5.0	2.0	2.0	9.0
	No. of patients	19.4	19.0	7.5	2.6	2.3	
U-BOIN	Selection %	11.9	74.1	3.7	0.4	0.0	9.9
	No. of patients	15.0	29.5	5.1	1.1	0.2	

Table 4.3 Continued:

		Dose Level					% of early
Design		1	2	3	4	5	stopping
	DLT probability	0.10	0.30	0.50	0.55	0.65	
	Efficacy probability	0.45	0.45	0.45	0.45	0.45	
	Utility	58.0	50.0	42.0	40.0	36.0	
EffTox	Selection %	69.0	27.0	2.0	0.0	0.0	2.0
	No. of patients	29.7	17.3	4.9	1.0	0.3	
U-BOIN	Selection %	75.4	22.8	1.5	0.2	0.0	0.2
	No. of patients	33.1	16.5	3.7	0.5	0.1	
Scenario 6							
	DLT probability	0.05	0.07	0.10	0.12	0.16	
	Efficacy probability	0.35	0.45	0.50	0.55	0.75	
	Utility	53.0	59.0	61.0	64.0	75.0	
EffTox	Selection %	10.0	12.0	26.0	24.0	29.0	0.0
	No. of patients	8.6	7.9	14.2	11.1	12.1	
U-BOIN	Selection %	5.9	11.7	13.1	13.6	55.7	0.0
	No. of patients	7.0	8.8	9.0	9.1	20.1	
Scenario 7							
	DLT probability	0.03	0.16	0.27	0.45	0.55	
	Efficacy probability	0.15	0.38	0.45	0.60	0.70	
	Utility	40.0	51.0	51.0	53.0	55.0	
EffTox	Selection %	1.0	33.0	54.0	9.0	1.0	2.0
	No. of patients	4.2	13.8	25.9	7.0	2.3	
U-BOIN	Selection %	2.0	45.0	41.0	10.0	1.0	1.0
	No. of patients	5.1	20.4	20.0	6.7	1.2	
Scenario 8							
	DLT probability	0.22	0.45	0.55	0.65	0.70	
	Efficacy probability	0.03	0.10	0.20	0.35	0.40	
	Utility	25.0	23.0	25.0	30.0	31.0	
EffTox	Selection %	1.0	6.0	6.0	0.0	0.0	87.0
	No. of patients	4.6	6.2	6.3	2.3	1.2	
U-BOIN	Selection %	0.8	5.5	1.7	0.0	0.0	92.0
	No. of patients	14.3	9.7	1.6	0.1	0.0	

Table 4.3 summarizes the operating characteristics for the designs. The U-BOIN design outperforms EffTox as it has a larger PCS and allocates more patients on the OBD. In scenarios 1 and 2, the dose-response curve increases first, and then plateaus, and the OBDs (dose level 2 and 3,

respectively) are one dose level lower than the MTD (i.e., dose level with DLT probability closest to the target DLT probability). The U-BOIN has a 23% higher PCS and assigns seven more patients on the OBD than EffTox does in scenario 1. Similarly, the U-BOIN has a larger PCS in scenario 2, for which the two designs have comparable patient allocation. In scenarios 3 and 4 where the dose-response curve increases to an optimal point and then decreases, U-BOIN has 22% and 27% higher PCS than EffTox, respectively. Moreover, U-BOIN assigns nine and ten more patients on the OBD, respectively in the two scenarios. In scenario 5 where the OBD is located on the first dose, and the response rate does not change with dose levels, U-BOIN has a 6% higher PCS and assigns three more patients on the OBD. In scenario 6, where all doses are safe and the OBD is the last dose level, the U-BOIN has a 27% higher PCS and assigns eight more patients on the OBD. This is because EffTox cannot distinguish well the suboptimal doses from the OBD in this scenario, subsequently assigning more patients on dose level 3 and 4 and selecting one of the dose levels as the OBD 55% of the time. U-BOIN has comparable performance to EffTox when there are two OBDs (scenario 7), in terms of PCS and patient allocation. When there is no OBD due to toxicity or futility (scenario 8), U-BOIN has a larger chance to stop the trial early.

4.3.2 Simulation B

Simulation B considers the case that Y_T is quickly ascertainable, but Y_E takes a long time to be scored with the assessment window of 3 months. We assume that patient accrual follows a Poisson process, with the rate of 3 patients per month. We simulated Y_T and Y_E based on the same Gumbel model and 8 scenarios, as described in Simulation A. The reason we chose the same simulation scenarios is that, by doing so, the results from Simulation A (i.e., Y_E is always observed) can be used as a benchmark to evaluate the performance of the design in Simulation B (i.e., Y_E is partially observed, due to delayed response). To generate a delayed response, for patients who experience efficacy in the assessment window (i.e., $Y_E = 1$), we simulate their time to efficacy from a truncated Weibull distribution with a support of $(0, 3)$ months. The shape and scale parameter for the Weibull distribution are chosen such that the efficacy probability at the end of assessment time matches those in Table 4.3, and that 90% of the responses occur in the latter half of the assessment window (i.e., $(1.5, 3)$ months). The immune response Y_I is generated from $N(\mu_{I,j}, 1)$, where $\mu_{I,j}$ is plotted in Figure 4.3. We set $C_I = 0.9$. Under each scenario, 2,000 simulations are performed.

Table 4.4 shows the simulation results. The PCS of the OBD and the number of patients

allocated to the OBD are generally comparable to the results in Simulation A (i.e., the optimal benchmark with fully observed data) and U-BOIN still outperforms the EffTox design. The results indicate that U-BOIN efficiently handles the delayed efficacy response. Because the U-BOIN design does not need to suspend accrual to wait Y_E to be fully observed and allows real-time decision making, it has great potential to shorten the trial duration.

Table 4.4: Results of Simulation B, including the selection percentage (selection %), the average number of patients treated at each dose (No. of patients), the percentage of early stopping, and the trial duration. The optimal biological dose (OBD) is bolded. In scenario 8, the OBD does not exist, and thus the percentage of early stopping is bolded.

		Dose Level					% of early	Duration
Design		1	2	3	4	5	stopping	(month)
Scenario 1								
EffTox	Selection %	2.0	50.0	45.0	2.0	0.0	0.0	45.9
	No. of patients	4.3	22.7	23.8	2.8	0.4		
U-BOIN	Selection %	0.9	68.9	24.1	2.9	0.3	2.8	20.8
	No. of patients	6.4	29.2	13.7	3.6	0.5		
Scenario 2								
EffTox	Selection %	0.0	4.0	60.0	29.0	7.0	0.0	43.8
	No. of patients	3.4	4.9	26.4	14.2	5.1		
U-BOIN	Selection %	0.0	0.2	64.1	22.4	4.9	8.3	20.0
	No. of patients	4.3	6.1	24.5	12.1	4.1		
Scenario 3								
EffTox	Selection %	12.0	70.0	10.0	1.0	0.0	7.0	45.6
	No. of patients	10.8	25.5	11.5	2.9	1.0		
U-BOIN	Selection %	0.8	89.6	6.0	0.8	0.0	2.8	20.7
	No. of patients	6.7	33.6	9.5	2.9	0.4	2.8	
Scenario 4								
EffTox	Selection %	36.0	47.0	5.0	2.0	2.0	9.0	47.8
	No. of patients	19.4	19.0	7.5	2.6	2.3		
U-BOIN	Selection %	6.8	70.4	3.4	0.4	0.0	19.1	18.8
	No. of patients	13.6	27.9	4.7	1.0	0.2	19.1	
Scenario 5								
EffTox	Selection %	69.0	27.0	2.0	0.0	0.0	2.0	50.0
	No. of patients	29.7	17.3	4.9	1.0	0.3		
U-BOIN	Selection %	72.9	22.7	2.2	0.2	0.0	2.0	20.7
	No. of patients	32.9	16.4	3.7	0.5	0.0	2.0	
Scenario 6								

Table 4.4 Continued:

		Dose Level					% of early	Duration
Design		1	2	3	4	5	stopping	(month)
EffTox	Selection %	10.0	12.0	26.0	24.0	29.0	0.0	43.4
	No. of patients	8.6	7.9	14.2	11.1	12.1		
U-BOIN	Selection %	4.2	10.3	12.2	13.6	59.2	0.4	21.0
	No. of patients	6.5	8.6	9.5	9.9	19.4	0.4	
Scenario 7								
EffTox	Selection %	33.0	45.0	45.0	43.0	40.0		46.5
	No. of patients	1.0	33.0	54.0	9.0	1.0	2.0	
U-BOIN	Selection %	5.0	45.6	33.1	10.5	1.7	4.1	20.7
	No. of patients	8.5	22.7	15.6	5.2	0.8	4.1	
Scenario 8								
EffTox	Selection %	1.0	6.0	6.0	0.0	0.0	87.00	16.6
	No. of patients	4.7	6.1	6.3	2.4	1.3		
U-BOIN	Selection %	0.0	0.4	0.4	0.0	0.0	99.10	8.8
	No. of patients	11.1	4.7	1.3	0.1	0.0		

To examine the performance of U-BOIN when there are delayed outcomes at a relatively smaller sample size, we conduct the simulation again using $N = 39$. Results show that U-BOIN still maintains its great operating characteristics, even when sample size is relatively small. The simulation results are provided in Table S8 of the Supplementary Material (at <https://onlinelibrary.wiley.com>). We also provide operating characteristics for eight additional representative scenarios (Scenarios A1-A8) for both Simulation A (Table S9) and Simulation B (Table S10). The results, again, show that U-BOIN has robust performance in various scenarios in comparison to EffTox.

4.3.3 Sensitivity analysis

We conduct sensitivity analyses to assess the robustness of the U-BOIN design by using a different set of utility values $\psi' = \{\psi_1 = 0, \psi_2 = 20, \psi_3 = 55, \psi_4 = 100\}$, which assigns a lower score for $(Y_E = 0, Y_T = 1)$ and a higher score for $(Y_E = 1, Y_T = 1)$, indicating that patients are willing to tolerate a higher toxicity risk in order to attain a higher efficacy. We also evaluate the performance of U-BOIN using different patient allocation strategies (i.e., step B2 of Stage II of the dose finding algorithm): PW, AR, and ER.

The simulation results (see Figure 4.4) show that the U-BOIN design performs well with comparable PCS and patient allocations for the two different utilities. This characteristic is important, since two clinicians might have slightly different opinions about the specific utility value assigned to an outcome combination. Our design shows robustness for this type of deviation.

Figure 4.5 shows the results under different patient allocation strategies. The PCS is comparable among the PW, AR, and ER strategies. The difference among the three strategies mainly lies in the number of patients treated on OBD. The PW approach, on average, assigns significantly more patients on the OBD than both the ER and AR approaches, while it has larger variability. Since the variation is more towards the higher end (i.e., more patients are treated on OBD), the PW approach is desirable in this simulation study.

Note that in this Simulation A, we specify $s_2 = N$ for fair comparison with the EffTox design, which does not stop the trial early on the basis of the number of patients treated on a dose. A sensitivity analysis using $s_2 = \{18, 21\}$ is provided in Section 5 of the Supplementary Material (at <https://onlinelibrary.wiley.com>). The result in Figure S1 shows that using the recommended values, the change in the percentage of correct selection is negligible, but the saving in sample size is substantial.

For the case with delayed Y_E , we performed three additional sensitivity analyses by (1) assuming that the efficacy assessment window is two months or four months; (2) specifying more vague prior distributions for the coefficients β_0 and β_1 with $\beta_0 \sim N(0, 3.75^2)$ and $\beta_1 \sim \text{Gamma}(\text{shape} = 1, \text{rate} = 0.4)$, both of which have standard deviations that are three times the previous values in section 4.2.5; and (3) considering two additional scenarios (scenarios B1 and B2 in Table S11), where Y_I and Y_E are weakly associated with Pearson's correlation coefficients -0.15 and 0.2, respectively. The results show that the U-BOIN design is also robust to the length of assessment window, prior distribution of the prediction model (Figure S2), and the association between Y_I and Y_E .

4.4 Summary

We proposed the U-BOIN, a seamless Phase I/II model-assisted design, to identify the OBD for targeted and immunotherapy trials. The U-BOIN design accounts for the efficacy-toxicity trade-off using a utility function. Unlike most existing Phase I/II designs that require complicated real-time model fitting and estimation to make dose assignment decisions, the U-BOIN is simple and easy to implement. The dose assignment rules of the U-BOIN can be pre-tabulated in decision tables and included in the trial protocol before the onset of a trial. To conduct the trial, no

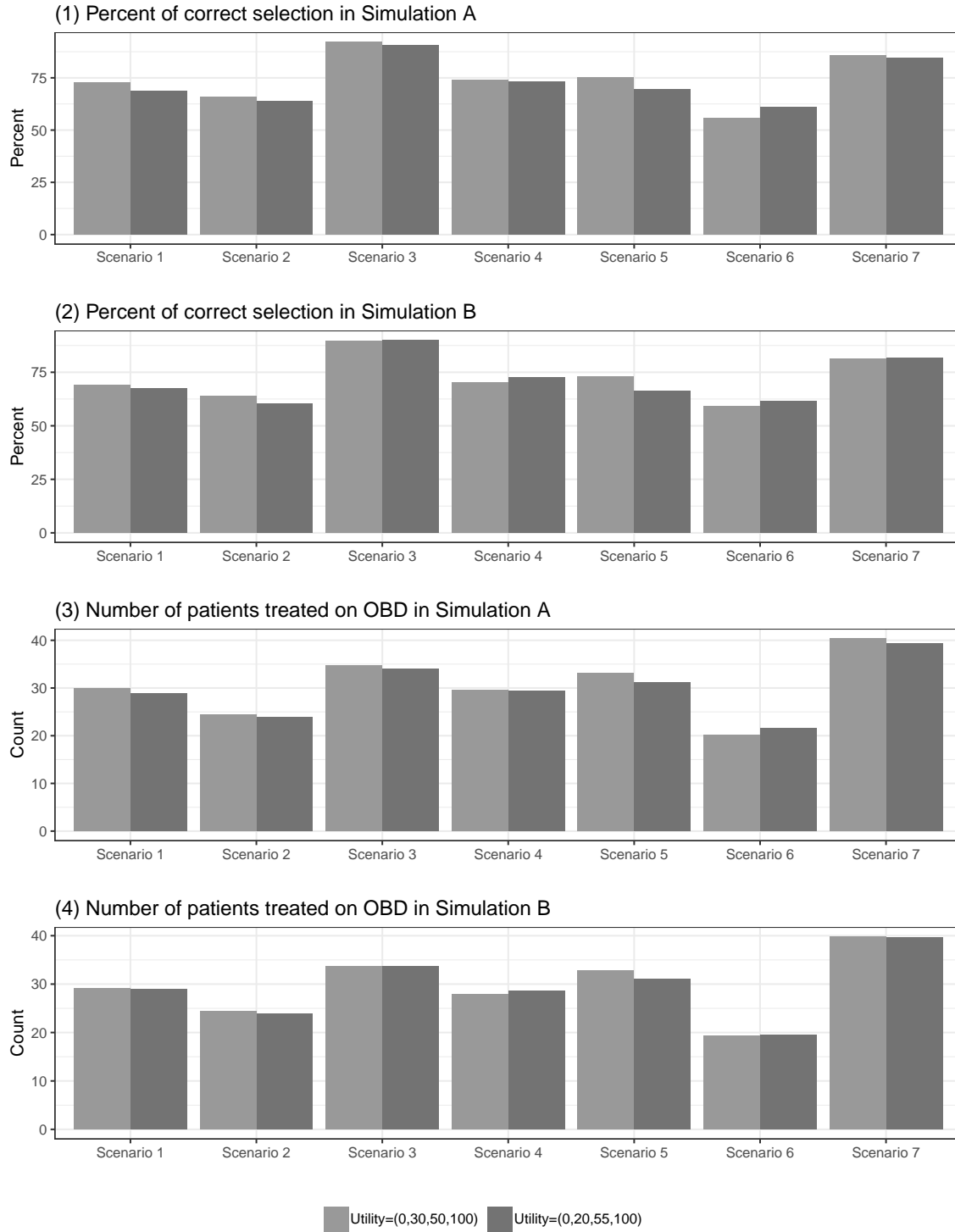


Figure 4.4: Results of sensitivity analysis for different utilities. Scenario 8 is not included, as the OBD does not exist in that scenario.

complicated calculation is needed. The investigator can simply use the decision tables to determine dose assignment. Simulation studies show that compared to a more complicated model-based Phase

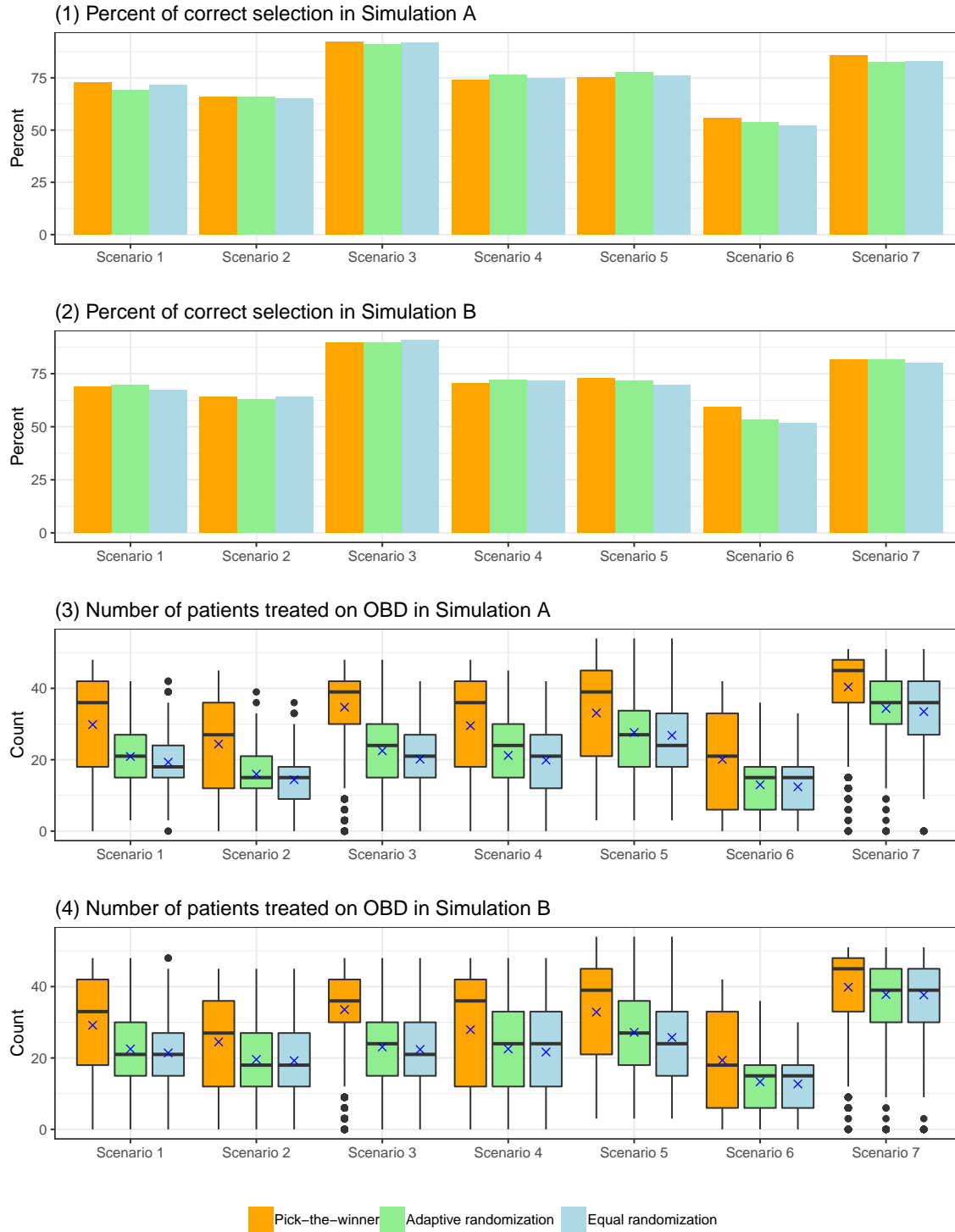


Figure 4.5: Results of sensitivity analysis for different patient allocation strategies: pick-the-winner (PW), adaptive randomization (AR), and equal randomization (ER). Scenario 8 is not included, as the OBD does not exist in that scenario.

I/II design, the U-BOIN has higher accuracy to identify OBD and is more robust. To facilitate the use of the U-BOIN design in clinical trials, we develop a user-friendly software freely available at www.trialdesign.org. Readers can find the U-BOIN design in the “*BOIN suite*” under Phase I applications.

While this project focuses on immunotherapy and targeted trials, U-BOIN also can be used for conventional cytotoxic agent trials. In such cases, both toxicity and efficacy typically increase with the dose, but may do so at different rates. It is likely that increasing the dose causes much higher toxicity, with limited efficacy gain. The idea of risk-benefit trade-off, and thus finding the optimal dose, is still generally applicable here for use in most medical decisions in practice.

Like most model-assisted designs, U-BOIN models efficacy and toxicity at each dose independently, whereas model-based Phase I/II designs (e.g., EffTox design) model efficacy and toxicity across all doses, through imposing a parametric dose-efficacy and -toxicity curve model. As a result, one may worry about the potential efficiency loss for U-BOIN. Our numerical study and previous studies show that, for the purpose of dose finding, the efficiency loss caused by using only local data (in model-assisted designs, such as U-BOIN) is minimal or negligible. This can be explained as follows. First, although U-BOIN models only the local data at the current dose, its dose-finding algorithm (e.g., escalate the dose if the current dose is safe, and de-escalate the dose if the current dose is toxic) implicitly uses the dose-toxicity order information across the doses. In addition, in practice, the parametric dose-efficacy and -toxicity model assumed by model-based designs is more likely to be miss-specified than correctly specified. Thus, on average, borrowing information across doses through the parametric model leads to rather limited efficiency gain. Furthermore, such limited efficiency gain does not necessarily translate into performance gain. This is because, to make correct decisions of dose assignment and selection, we only need to correctly estimate the rank of utility across the doses. A slightly more variability on the estimate of the utility has no or negligible impact on the performance of the design.

A Appendix

Theorem 1. Marginal-probability-based trade-off U_j^M , defined in (4.3), is a special case of the utility method defined in (4.1), in the sense that for a prespecified weight w , we can find (ψ_2, ψ_3) , such that $U_j = \xi U_j^M$, where ξ is a non-zero constant.

Proof If there is a constant ξ , such that $U_j^M = \xi \tilde{U}_j$, then, \tilde{U}_j is a special case of U_j^M .

To fix scale for utility, we propose that $\psi_1 = 0, \psi_4 = 100$. Then U_j^M is a function of only ψ_2 and ψ_3 .

$$U_j^M = \psi_2 \pi_{j2} + \psi_3 \pi_{j3} + 100 \pi_{j4}$$

It is obvious that $U_j^M > 0$ by definition. Denote $\tilde{U}_j(w) = \tilde{U}_j = \pi_{E,j} - w \pi_{T,j}$. If $U_j^M = \xi \tilde{U}_j$, then $\xi \neq 0$, $\tilde{U}_j(w) \neq 0$, and $\psi_2 \pi_{j2} + \psi_3 \pi_{j3} + 100 \pi_{j4} = \xi \tilde{U}_j(w)$. Since $0 < \psi_2 \pi_{j2} + \psi_3 \pi_{j3} < 100(\pi_{j2} + \pi_{j3})$, we have $\frac{100 \pi_{j4}}{\tilde{U}_j(w)} < \xi < \frac{100(1 - \pi_{j1})}{\tilde{U}_j(w)}$, which is nonzero constant.

In summary, for a prespecified w and $\tilde{U}_j(w) \neq 0$, we can find (ψ_2, ψ_3) and a non-zero constant ξ , such that $U_j^M = \xi \tilde{U}_j$, which says that \tilde{U}_j is a special case of U_j .

CHAPTER 5

Versatile software platforms for the implementation of novel model-assisted designs in early phase clinical trials

5.1 Introduction

The main purpose of Phase I studies is to evaluate the safety of a new drug and identify a maximum tolerated dose (MTD), defined as the highest dose with acceptable level of toxicity. Typically, the primary endpoint in Phase I trial is binary, i.e., whether or not a patient experiences dose-limiting toxicity (DLT) while receiving a dose in the trial. The corresponding parameter of interest is the proportion of patients who experience DLTs (i.e., DLT probability). The designs used to identify the MTD can be classified into three types according to their statistically foundations and implementation approaches: algorithm-based (also known as rule-based), model-based, and model-assisted [115]. Algorithm-based designs guide dose assignment through a set of prespecified rules without assuming any parametric assumption on the dose-toxicity curve. Thus, they are simple to implement. But the downside is that they are rigid and have poor accuracy [117]. A classical example is the 3+3 design [19]. Model-based designs often assume a dose-toxicity model and guide the dose escalation and de-escalation by continuously updating the estimate of dose-toxicity model based on the accumulated data. Typical examples of the model-based designs are the continual reassessment method (CRM [77]) and its variations, such as escalation with overdose control (EWOC[2]), Bayesian logistic regression model (BLRM [74]), and Bayesian model averaging CRM (BMA-CRM [111]). Model-based designs yield better performance than the 3+3 design in identifying the MTD and allocating more patients to the MTD [124, 125]. However, model-based designs are not straightforward for clinicians to comprehend, which has limited their use in practice.

Model-assisted designs are a class of recent designs that enjoy both the simplicity of

algorithm-based designs and the satisfactory performance of model-based designs. Similar to model-based designs, model-assisted designs use a statistical model (e.g., the binomial model for the number of DLTs) to assist the derivation of the design for efficient decision making; and like algorithm-based designs, model-assisted designs can pre-tabulate the dose escalation and de-escalation rules before the onset of a trial, making it as simple as the 3+3 to implement. Classic examples of model-assisted designs in Phase I clinical trials include the Bayesian optimal interval (BOIN) design [65], keyboard design [105], and modified DLT probability interval design (mTPI [42]) and its extension mTPI-2 [31].

Zhou et al. [124, 125] compared the BOIN design to other model-assisted and model-based designs. They found that BOIN and keyboard designs both perform better than 3+3 in terms of accuracy, safety, and reliability, and are safer than CRM and mTPI. We also compared BOIN to i3+3, and further shows its superiority over algorithm-based designs in Chapter 2.1. The BOIN and keyboard designs have been extended to accommodate commonly encountered challenges in Phase I clinical trials, such as fast accrual, delayed toxicity, and evaluation of drug combinations. There is a well-validated *BOIN* R package to implement the BOIN designs. But using R package requires a certain level of statistical programming. Thus, it remains a challenge for practitioners to evaluate the operating characteristics of the designs in a timely manner without programming skills. There is no R package for the keyboard design and its extensions, making it even harder for non-statisticians to evaluate the accuracy and reliability of the designs. Despite the well-documented flaws of 3+3, it is still the dominant method used in Phase I clinical trials. The lack of reliable, robust, and easy-to-use software has limited the use of the more superior novel designs such as BOIN and Keyboard. Lee and Chu [39] suggested that “the availability of accompanying software for the implementation of Bayesian methods is crucial for the use of these methods in clinical trials”.

The R package *Shiny* has become a popular tool for software development. It enables R users to develop interactive web applications directly in R [10], and has revolutionized the way statisticians share their analytic results and research methods. Applications developed using *Shiny*, referred to as shiny apps hereafter, typically have interactive point-and-click interface that are self-explanatory and accessible to a broader audience. It can empower people without statistical programming skills to explore the performance of a novel design easily without knowing statistical programming. For this reason, shiny apps have gain a lot of popularity in both industry and academia. Many industry companies, such as *Novartis* and *Eli Lilly and Company*, have a team specializing in analyzing, visualizing, and reporting clinical trial data using Shiny apps. In academia,

shiny apps have been used to facilitate the understanding and use of novel methods. Examples include the web applications for model-based dose escalation/de-escalation design [79, 102] and for planning and simulating adaptive platform trials [100], among others. To facilitate the use of model-assisted designs for designing and conducting Phase I clinical trials and make it more accessible for clinicians who may not have expertise in statistical programming, we develop a software platform for BOIN designs (BOIN suite) and keyboard designs (Keyboard suite), respectively.

One distinct feature of our software is that each of the two suites includes multiple shiny apps for model-assisted designs, each tackling a different challenge in Phase I clinical trials. BOIN suite include apps for single-agent trials using non-informative prior (standard BOIN [65]), incorporating historical data (iBOIN [127]), and having fast accrual or delayed toxicity (TITE-BOIN [116]); for drug-combination (BOIN COMB) to find single MTD [56] or MTD contour [121]; and for utility-based single-agent Bayesian optimal interval design (U-BOIN [126]) to identify optimal biological dose (OBD). Keyboard suite includes apps for single-agent trials without delayed toxicity (KEYBOARD) [105] or with delayed toxicity (TITE-KEYBOARD [57]); and for combination trials to find single MTD (KEYBOARD COMB [80]). Another unique feature of the two software suite is that the designs considered are all model-assisted and thus provide simple-to-use decision rules like those in 3+3, which is easy for practitioners familiar with 3+3 to adopt these new adaptive designs.

Functionality-wise, all the shiny apps have a simple and user-friendly interface to carry out the following tasks: (1) generate decision tables and design flowchart; (2) simulate operating characteristics such as probability of selecting the MTD (or OBD), percentage (or number) of patients allocated to each dose, and early stopping probability; (3) prepare trial protocol; (4) compute the recommended dose level for the next cohort of patients, based on data obtained; and (5) estimate MTD (or OBD) when a study is completed.

The greatest contribution of this work is that, the software platform empowers clinicians and even statisticians to easily evaluate the performance of the designs and aid them to choose appropriate designs for their intended trials. Through simulation information on the interactive shiny apps, clinicians can evaluate the accuracy and safety of the novel designs in a timely manner. That is, they can change the trial specifications and get real time feedback on how the designs would perform in various trial settings. The availability of these software will not only make the designs easier to access and enable a more efficient collaboration between statisticians and clinicians, but will also have a profound impact on early phase drug development, as the availability of these shiny

apps will substantially increase the use of Bayesian adaptive designs in early phase clinical trials.

5.2 Method

The software platform we developed for Phase I clinical trial design consists of two versatile suites: BOIN suite and Keyboard suite. We first describe the designs under each suite, respectively, and then provide information on the statistical tools used to develop the two software suites.

5.2.1 BOIN suite

The BOIN suite includes shiny apps for single-agent trials using non-informative prior (the standard BOIN) or informative prior (iBOIN); single-agent trials with fast accrual or delayed toxicity (TITE-BOIN); combination trials (BOIN COMB) to find a single MTD or the MTD contour; and utility-based Bayesian optimal interval design (U-BOIN) to identify the OBD.

Standard BOIN design for single-agent trials

The BOIN design was constructed to minimize the chance of making incorrect dose assignment decisions in order to effectively treat patients, i.e., minimizing the probability of exposing patients to overly toxic or sub-therapeutic doses. The implementation of the design can be done by simply comparing the observed DLT probability to a pair of optimal escalation and de-escalation boundaries (λ_e, λ_d) , which are derived given the specification of target DLT probability (derivation can be found in Liu and Yuan [65]). Let $\hat{p}_j = y_j/n_j$ denote the observed dose limiting toxicity (DLT) probability at dose $j = 1, \dots, J$, where y_j is the number of observed DTLs and n_j is the number of patients treated at dose j . A trial using BOIN can be conducted as simple as below. Suppose the current dose is j , then the dose used for next cohort of patients is determined based on the following rule.

- escalate to dose j , if $\hat{p}_j \leq \lambda_e$ and $j \neq J$;
- de-escalate to dose $j - 1$, if $\hat{p}_j \geq \lambda_d$ and $j \neq 1$
- stay at the current dose, otherwise.

The decision rule is similar to that in the cumulative cohort design (CCD [37]), which was developed based on the asymptotic distribution of patient allocation of the group up-and-down design [28] using the Markov chain theory [57]. In comparison, the advantage of BOIN is that it has better finite sample performance, which is desirable in Phase I trials given the typically small sample size.

For the purpose of overdose control, doses j and higher levels will be eliminated from further examination if $\Pr(p_j > \phi \mid \text{data}) > \delta$ and at least three patients are treated on dose j , where p_j is the true DLT probability of dose level $j, j = 1, \dots, J$, ϕ is the target DLT probability, and δ is a probability cutoff (typical values are 0.9 and 0.95). This posterior probability is evaluated based on the beta-binomial model $y_j \mid p_j \sim \text{binomial}(n_j, p_j)$ with $p_j \sim \text{uniform}(0, 1)$. When the lowest dose is eliminated, stop the trial for safety.

The BOIN app tabulates decision rules using the aforementioned algorithm and makes the implementation of the design as simple as the traditional 3+3. Suppose the target DLT probability is 0.3, with the default setting of the BOIN design, the shiny app will produce the decision table as shown Table 5.1. Thus, conducting a trial simply requires comparing the number of observed DLTs at the current dose to the boundaries provided in the table.

Table 5.1: Dose escalation/de-escalation and elimination boundaries in BOIN design with target DLT probability of 0.3

	Number of patients treated at current dose											
	1	2	3	4	5	6	7	8	9	10	11	12
Escalate if # of DLT \leq	0	0	0	0	1	1	1	1	2	2	2	2
Deescalate if # of DLT \geq	1	1	2	2	2	3	3	3	4	4	4	5
Eliminate if # of DLT \geq	NA	NA	3	3	4	4	5	5	5	6	6	7

Note. The NA's in the third row means that the elimination rule will not be triggered when the number of patients treated on current dose is less than three.

In addition to the decision boundaries, the shiny app also provides design flowchart and simulation results for use in protocol preparation, upon users' input of trial parameters. To minimize the workload of trial planning for statisticians and clinicians, the app is also equipped to output well-organized template that includes a brief description of the design, followed by the design flowchart, decision tables, and simulation results. To enable a fast comparison between BOIN and 3+3 designs, the app also provides the option for users to select whether the 3+3 design is simulated while running simulation for the BOIN design (a concrete example for using this app is provided in section 5.3.2).

BOIN for single-agent trials using informative prior (iBOIN)

Model-assisted designs (e.g., BOIN) were developed assuming non-informative prior distribution on the parameter of interest. Little research has been done regarding to how informative prior information that is obtained from historical trials or real-world evidence can be incorporated into the model-assisted designs without compromising their simplicity for trial implementation.

This has led to the misconception that model-assisted designs cannot incorporate informative prior information, which is sometimes cited as their weakness, compared to model-based designs. We recently proposed a unified framework to incorporate informative prior information into model-assisted designs, including BOIN and keyboard/mTPI-2 designs (details provided in Chapter 3). Numerical studies show that incorporating appropriate prior information can improve the performance of the model-assisted designs in a similar way as in the CRM. We found that BOIN with informative (referred to as iBOIN) has the most desirable accuracy and reliability. Specifically, iBOIN utilizes the skeleton approach same as in the CRM, combined with the concept of prior effective sample size [69]. The dose assignment is based on similar algorithm described for the standard BOIN, except that by incorporating informative prior, the escalation and de-escalation boundaries (λ_e, λ_d) in iBOIN vary with the dose levels and the number of patients treated at the doses. However, the iBOIN app can provide decision table similar to that in Table 5.1 for trial implement, maintaining the simplicity feature of the BOIN design. Numerical studies have shown that when prior information of good quality is available, the iBOIN can greatly increase the accuracy to identify MTD and allocates more patients at the MTD.

The iBOIN app requires similar input as the standard BOIN, except that it also requires the specifications of prior DLT probability and prior effective sample size for all doses in the trial. Additionally, it has the flexibility for users to determine if the prior DLT probabilities are used to select the MTD at the end of the trial. The app also has the options for simulation and protocol preparation.

BOIN for single-agent trial with delayed toxicity (TITE-BOIN)

The typically long duration of drug development has posed great challenges in meeting the ever increasing need of effective therapies. To speed up drug development, increasing the pace of patient accrual and using novel adaptive designs have become the two most common strategies [56, 115, 116]. BOIN design is one of the novel adaptive designs and has been validated and used in many recent trials [115], but it is not a panacea. BOIN requires the toxicity outcomes to be quickly ascertainable in order to have complete observed data for decision making at an interim time. With the development of novel molecularly targeted agents and immunotherapy, late-onset (delayed) toxicity has become a common problem encountered in Phase I trials. A trial with delayed toxicity outcome causes major logistic difficulty when using existing adaptive Phase I trial designs that require quick observance of toxicity to determine the dose assignment for next cohort of patients.

The same logistic difficulty arises when the accrual is faster than the toxicity evaluation window. To address these challenges, Yuan et al. [116] proposed the time-to-event Bayesian optimal interval (TITE-BOIN) design that allows real-time dose assignment decisions for new patients despite the existence of pending toxicity data. The essential idea of TITE-BOIN is to impute the DLT outcome for patients whose DLT data are pending, by using data from all patients and their follow-up times in a time-to-event model. The TITE-BOIN has comparable accuracy to identify the MTD, when compared to the more complex model-based time-to-event continuous reassessment method (TITE-CRM [14]), but it is simpler to implement with substantially better overdose control.

TITE-BOIN can also pre-tabulates dose-escalation/de-escalation rules before the onset of a trial. To conduct a trial, it requires only the number of patients treated on each dose, the corresponding number of patients who experience DLT, the number of patients whose DLT data is pending, and the standard total follow-up times (STFT), which is determined using the sum of the follow-up times for all pending patients at the current dose divided by the DLT assessment window. The TITE-BOIN app provides three different time-to-event models for predicting pending toxicities. Users can choose any of them to determine the decision table for trial conduct. It also has a built-in calculation to calculate the STFT value to assist in the use of the decision table. Users can also use the app to assess the accuracy and safety of the design through its built in simulation functions. Protocol templates can be downloaded for trial planning as well. Thus, compared to TITE-CRM [14] that requires repeat model fitting during the trial, the TITE-BOIN is more transparent and also more straightforward to use with the availability of the shiny app. The rolling six (R6) design [90], an algorithm-based design, has often used to address the logistic difficulties caused by either fast accrual or delayed toxicity. The TITE-BOIN app also provides an option for users to determine if they want to include R6 design in the simulation.

BOIN for drug-combination trials (BOIN COMB)

The traditional paradigm that one drug for one disease may fail in circumstances where more than one pathways drive the malignant process of the tumor under investigation [78]. Designing combination trials could be the most desirable way to pursue greater clinical benefits for patients in this case. It is more difficult to design drug-combination than single-agent trials due to two main challenges: (1) the drug combinations are only partially ordered in terms of their toxicity probabilities, and (2) there could exist a MTD contour (i.e., more than one MTD). While there are numerous model-based Bayesian Phase I trials for combination trials, their use is limited

due to the complexity of their statistical foundation and computation as well as their sensitivity to prior specifications. To address the first challenge, Lin and Yin [56] extended the BOIN design that addresses one-dimensional dose space to explore the two-dimensional dose space for MTD identification. Zhang and Yuan [121] developed the waterfall design, which is also built upon the BOIN design, to simultaneously address both of the two challenges aforementioned. The two designs also belong to the model-assisted class and provide simple and robust alternatives for Phase I combination trials. An overview of the two designs are provided below.

(1) BOIN for finding a single MTD in drug-combination trials

Suppose there are two drugs (A and B) under investigation. Drug A has J dose levels while B has K dose levels. Let p_{jk} denote the true DLT probability of the combination doses (j, k) , where $j, j = 1, \dots, J$ and $k, k = 1, \dots, K$. And denote the corresponding observed DLT probability $\hat{p}_{jk} = y_{jk}/n_{jk}$, where n_{jk} is the number of patients treated on the dose combination (j, k) and y_{jk} is the number of patients who experience DLT. The design proposed by Lin and Yin [56] makes the decision of dose escalation/de-escalation based on a similar rule as the single-agent BOIN design described above. To implement the trial, it just needs to compare the observed DLT probability to a pair of escalation and de-escalation boundaries: (λ_e, λ_d) , which are optimized by minimizing the probability of incorrect dose assignment (see Lin and Yin [56] for more details).

The difference between single-agent BOIN and the design proposed by Lin and Yin [56] for combination trials is that there is more than one neighboring doses to move to when escalation (when $\hat{p}_{jk} \leq \lambda_e$) or de-escalation (when $\hat{p}_{jk} \geq \lambda_d$) decision is made. Suppose the current dose combination is (j, k) , to determine the assignment of the dose for next cohort of patients, the design first defines an admissible escalation set $A_E = \{(j+1, k), (j, k+1)\}$ and de-escalation set $A_D = \{(j-1, k), (j, k-1)\}$, and then decide which dose to use for next cohort of patients based on the value of p_{AT} , which is the posterior probability that the DLT probability of a dose combination (j', k') falls within the acceptable toxicity interval (λ_e, λ_d) , where (j', k') belongs to the escalation set A_E or the de-escalation set A_D . The dose assignment proceeds as follows given the observed data.

- Escalate to the dose combination in A_E that has the largest value of p_{AT} , if $\hat{p}_{jk} \leq \lambda_e$;
- De-escalate to the dose combination in A_D that has the largest value of p_{AT} , if $\hat{p}_{jk} \geq \lambda_d$;
- Stay at current dose combination, otherwise.

(2) BOIN for finding the MTD contour in drug-combination trial

Due to drug-drug interaction, it is of intrinsic interest to find multiple MTDs for many drug-combination. The MTDs can be further evaluated in subsequent Phase II trials to identify the one with the highest efficacy. To adequately explore the two-dimensional dose space requires a large sample size, because the toxicity probabilities in the dose space is partially ordered. This poses a great challenge in Phase I combination trials because Phase I trials typically have small sample size. To overcome the problem, Zhang and Yuan [121] proposed the waterfall design. The basic idea of the waterfall design is straightforward: divide the two-dimensional dose-finding problem into a series of simpler one-dimensional dose-finding subtrials. In each of the subtrials, dose toxicity probabilities are fully ordered and the standard BOIN is used to identify the MTD in each of the subtrials. The innovation of the waterfall designs is that the subtrials are conducted in a certain order such that the precedent subtrial will inform the conduct of the subsequent trial. The sequence of the subtrials efficiently borrows information across subtrials and thus saves sample size.

Suppose that the two drugs A and B has 3 and 5 dose levels, respectively. Figure 5.1 provides an example to illustrate the waterfall design. As shown, the waterfall design divides the 3×5 dose space into three subtrials. As shown in panel (a), the trial starts by conducting the first subtrial with the starting dose (1,1). After the first subtrial identifies (3,1) as the candidate MTD, waterfall design proceeds to the second subtrial with the starting dose (2,2) as shown in panel (b). After the second subtrial identifies (2,4) as the candidate MTD, the design conducts the third subtrial with the starting dose (1,5) (see panel (c)). After all the subtrials are completed, select the MTD contour based on the data from all the subtrials (panel (d)) using matrix isotonic regression [7].

The BOIN COMB shiny app gives users the option to select if the trial aims at finding a single MTD or a MTD contour. For either option, a decision table (similar to Table 5.1) for implementing the designs will be provided for trial conduct. If the purpose is to find the MTD contour, the app also provides the design chart (as shown in Figure 5.1) given the number of dose levels for drug A and drug B, and the starting dose combination of the two drugs. Like BOIN, iBOIN, and TITE-BOIN, the BOIN COMB app also have functions to conduct simulation and download trial protocol.

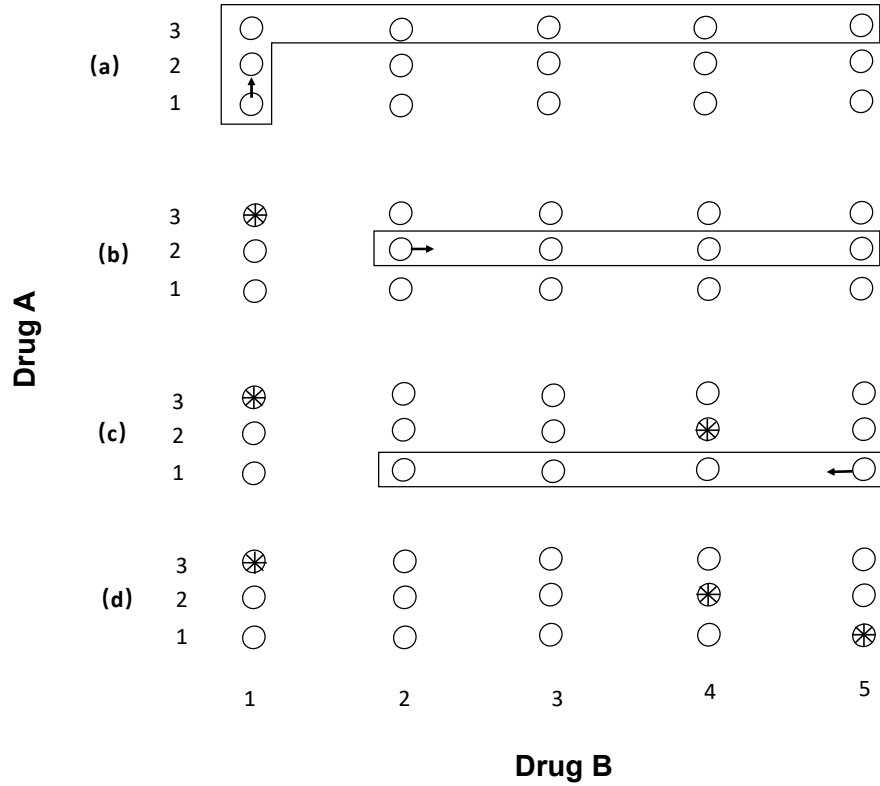


Figure 5.1: The waterfall design for finding the MTD contour in drug-combination trials (adapted from Figure 1 in Zhang and Yuan [121]). For panels (a)-(c), the circles surrounded by a rectangle is a subtrial. In panels (b)-(c), the filled circle is the MTD candidate selected from the previous subtrial. The MTD candidate for the third subtrial in panel (c) is indicated in the third row of panel (d).

Utility-based BOIN design (U-BOIN) to identify optimal biological dose

Traditional dose-finding designs for chemotherapy aim to find the maximum tolerated dose (MTD). The underlying assumption is that both toxicity and efficacy monotonically increase with the dose, and thus the MTD presents the most efficacious dose that is safe. This assumption, however, is often questionable for immunotherapy and targeted agents. Although it is reasonable to assume that toxicity increases with the dose, the same is not necessarily true for efficacy. To optimize the treatment benefit of immunotherapy and targeted therapy, therefore, it is important to consider toxicity and efficacy simultaneously and their risk-benefit trade-off during dose finding. The objective of dose finding for targeted therapy and immunotherapy is to identify the optimal biological dose (OBD), defined as the dose that has the highest desirability in terms of the risk-benefit trade-off.

Zhou et al. [126] developed the U-BOIN design, a utility-based seamless Phase I/II design

to find the OBD. In stage I, the BOIN design [65] is used to quickly explore the dose space and collect preliminary toxicity and efficacy data. In stage II, in light of accumulating efficacy and toxicity from both stages I and II, U-BOIN jointly models toxicity and efficacy using a multinomial-Dirichlet model and employs a utility function to measure dose risk-benefit trade-off. Dose assignment is guided using the contiguously updated posterior estimate of the utility for each dose after each cohort. To accommodate the delayed efficacy observed in some targeted and immunotherapy trials, the U-BOIN is capable of predicting the delayed efficacy using a short-term endpoint (e.g., immune activity or other biological activity of targeted agents) to facilitate real-time decision making. Extensive simulation studies showed that U-BOIN is not only simple to use with its pre-tabulated decision tables, but also has superior performance than model-based Phase I/II designs (such as the EffTox design [96]).

With the U-BOIN app, users can simply input trial parameters to obtain readily available decision boundaries to implement the trial for stage I and stage II. The app also provides a trial conduct function to determine dose assignment if users prefer using the app over decision tables. To use this function, users just need to upload a file with observed patient outcomes according to a prespecified format, which is provided in the app. Then, the app will automatically determine which dose to use for next cohort of patients. Like any other apps in the BOIN suite, the U-BOIN app can be used to run simulations and prepare trial protocol.

5.2.2 Keyboard suite

The Keyboard suite includes shiny apps for single-agent trials without delayed toxicity (KEYBOARD) or with delayed toxicity (TITE-KEYBOARD), and drug-combination trials (KEYBOARD COMB). The three model-assisted designs have comparable performance to their counterpart BOIN designs: standard BOIN, TITE-BOIN, and BOIN COMB. The difference is that the decision rules derived in the keyboard designs are based on different statistical approach. Although keyboard designs are considered not as transparent as BOIN designs [115], they still provide good alternatives for conducting Phase I trials. Thus, it is worthwhile to provide user-friendly applications.

Keyboard for single-agent trials

Keyboard design was developed to overcome the limitation of the modified toxicity interval (mTPI) design that is likely to overdose patients in a trial. The dose assignment in keyboard relies

on a prespecified proper dosing interval (referred to as the *target key*) and the calculation of a *strongest key* based on the posterior distribution of the DLT probability. The posterior distribution is determined by a beta-binomial model with uniform(0,1) prior distribution.

$$\begin{aligned}
 (5.1) \quad & y_j \mid n_j, p_j \sim \text{Binom}(n_j, p_j) \\
 & p_j \sim \text{Beta}(1, 1) \\
 & p_j \mid D_j = (y_j, n_j) \sim \text{Beta}(y_j + 1, n_j - y_j + 1)
 \end{aligned}$$

The *strongest key* refers to a probability interval that has the same width as the target key and that also has the largest posterior probability density based on accumulated data on current dose. Specifically, suppose the target DLT probability is 0.25 and (0.2, 0.3) is considered as the proper dosing interval (the target key). Then the probability intervals (0, 0.2) and (0.3, 1) on the two sides of the target key are divided into intervals with length 0.1, equal to that of the target key. As shown in Figure 5.2, this ends up with two and seven keys on the left and right sides of the target key, respectively. Suppose the current dose is j , the dose assignment for next cohort of patients is determined by comparing the location of the strongest key and the target key.

- Escalate to $j + 1$, if the strongest key is on the left side of target key;
- De-escalate to $j - 1$, if the strongest key is on the right side of the target key;
- Stay at the current dose, if the strongest key overlaps with the target key.

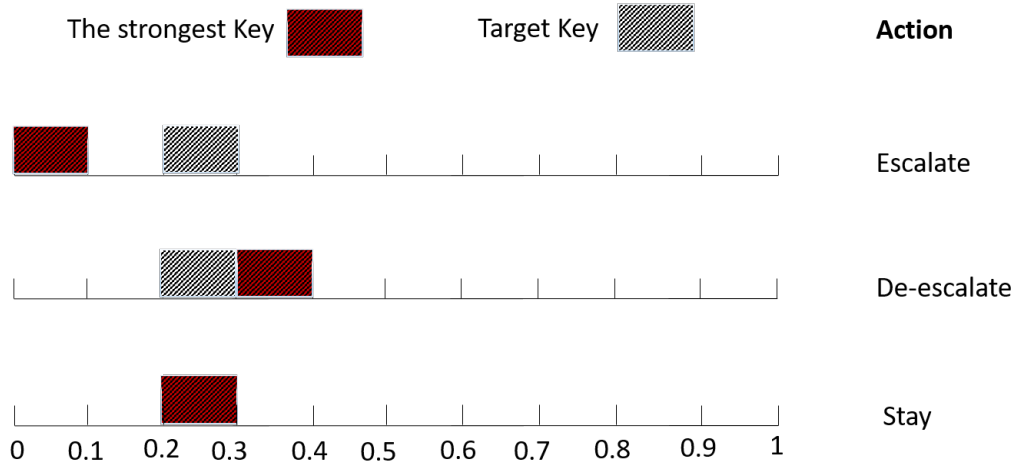


Figure 5.2: Illustration of dose escalation and de-escalation in keyboard design

With the shiny app, users do not need to calculate the posterior probability and the strongest key. Instead, the app will do all the calculations behind the scene and output a de-

cision table similar to that for BOIN (e.g., Table 5.1). This greatly simplifies the implementation of the trial design. The app can also be used to evaluate the accuracy and safety of the keyboard design through simulations and to prepare statistical protocols for trial planning.

Keyboard for single-agent trial with delayed toxicity (TITE-KEYBOARD)

The keyboard design requires the calculation of posterior distribution of toxicity rate at each dose that has been used to treat patients. The derivation of the posterior distribution requires fully observed toxicity outcomes for all patients enrolled in the trial as shown in (5.1). When fast accrual or delayed toxicity occurs, the likelihood that used to derive the posterior becomes a complicated product of the likelihood for patients who have finished the follow-up time (i.e., the toxicity evaluation window) with their toxicity outcomes evaluated and that for those whose toxicity outcome is still pending. This leads to a posterior distribution whose inference depends on Markov Chain Monte Carlo simulation and makes the decision rules impossible to tabulate before the onset of a trial. This feature loses the most important property of the model-assisted design: simplicity. To circumvent the issue, Lin and Yuan [57] proposed the TITE-keyboard, which uses an approximated likelihood for patients whose toxicity outcomes are pending, i.e., the toxicity outcomes for these patients have not been observed at the decision time. With the approximation, the joint likelihood for all the patients in the trial becomes the standard binomial likelihood. Specifically, to determine the posterior distribution of DLT probability, in the beta-binomial model (5.1), n_j is replaced by $ESS_j = y_j + \tilde{m}_j$, where ESS_j is known as the effective sample size at dose j , and \tilde{m}_j is the effective number of patients who have not experienced DLT. The value of \tilde{m}_j is determined as below.

$$\tilde{m}_j = \text{The number of patients who have completed the assessment without experiencing DLT at dose } j \\ + \frac{\text{Total follow-up time for patients with pending toxicity outcome at dose } j}{\text{Length of assessment window}}$$

For patient safety, TITE-keyboard does not permit dose escalation unless at least two patients have completed toxicity assessment at the current dose level. Thus, to conduct the trial with the TITE-Keyboard, no real-time model fitting is needed. Investigators only need to count the number of patients treated at current dose, the number of patients who have completed toxicity assessment, the number of patients with pending outcome, and calculate the effective sample size (ESS) at current dose. The TITE-KEYBOARD app not only provides a calculator for determining the ESS, but also returns a decision table to ease the implementation of the design. Similar to

other apps in the software platform, TITE-KEYBOARD app can also be used for simulation and protocol preparation.

Keyboard for drug-combination trials (KEYBOARD COMB)

The keyboard combination design [80] provides an alternative for drug-combination trials, as it has comparable operating characteristics to the BOIN combination design. The main difference between Keyboard combination design and BOIN combination design is the statistical principle used to determine the rules for dose escalation and de-escalation. The keyboard combination design relies on the calculation of posterior distribution for the DLT probability whereas the BOIN combination directly uses observed DLT probability.

There are five escalation and de-escalation schemes proposed in Pan et al. [80]. We only include the escalation and de-escalation scheme that is similar to that of BOIN combination as it yields the best operating characteristics under various scenario setting in the simulation studies in the original paper. Specifically, once escalation (or de-escalation) decision is made, keyboard combination determine the dose for next cohort of patients as the one with the largest posterior probability falling into the proper dosing interval (λ_e, λ_d) in the admissible escalation set A_E (or de-escalation set A_D). Unlike the BOIN combination design where the values (λ_e, λ_d) are determined by minimizing incorrect decisions, the keyboard combination designs use the values directly elicited from clinicians. Pan et al. also established the statistical properties of the Keyboard combination design such as optimality, coherence, and consistence, which assure clinicians that the design for sure will yield desirable operating characteristics.

The Keyboard combination design maintains the simplicity of the original single-agent keyboard design, which can tabulate decision rules before the onset of a trial. The KEYBOARD COMB app requires the number of dose levels for the drugs and their respective starting dose levels, as well as the target DLT rate and the proper dosing interval. All the other settings are similar to that in BOIN COMB app. With the input, the app will provide decision tables for trial conduct. Users can use the app to conduct simulation studies and prepare trial protocols.

5.2.3 Tools for software development

The BOIN suite and Keyboard suite were developed using R programming [95], primarily with the R *shiny* package [10]. Additional packages were used to enhance the appearance and functionality of the software. We used *ShinyBS* to add pop-up windows to show scenarios uploaded

or to save scenarios entered; *Shinyjs* to improve user experience of shiny apps by hiding, disabling, or enabling certain input elements as necessary; *rhandsonable* to make complicated trial settings (such as two-dimensional dose toxicity probabilities) in nice and concise table format; *DT* to output simulation results in HTML tables that can be saved in the formats of pdf, csv, excel, etc; *XLConnect* and *XLConnectJars* to write Microsoft Excel files that can be downloaded from the app; and *rmarkdown* to produce protocol template.

5.3 Results

5.3.1 Features of the applications

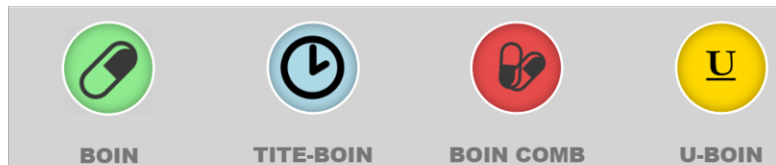
Different from most published shiny applications that only contain a single app for a specific model-based design, the software platform we developed includes two classes of model-assisted Bayesian adaptive designs: the Bayesian optimal interval (BOIN) suite and the Keyboard suite, shown in Figure 5.3 panels A and B, respectively. The BOIN suite contains BOIN design for single-agent trials with non-informative prior (standard BOIN) and informative prior (iBOIN); for single-agent trials with fast accrual or delayed toxicity (TITE-BOIN); for single-agent trials to identify optimal biological dose using utility-based approach (U-BOIN), and for drug-combination trials (BOIN COMB). Similarly, the Keyboard suite contains designs for single-agent trials without delayed toxicity (standard Keyboard) and with delayed toxicity (TITE-KEYBOARD), and for drug-combination (KEYBOARD COMB).

For the BOIN suite, a click on the icon for “BOIN” will bring out a sub-panel containing standard BOIN and iBOIN apps. All the app interfaces in the two suites are user-friendly and trial inputs are self-explanatory. Users do not need to have statistical programming skills to evaluate the operating characteristics of the trial designs. The evaluations can be done in a point-and-click fashion. Extensive help files are also available in each app. Panel C in Figure 5.3 provides a decision tree to help users determine which design is appropriate under specific trial settings in case they are not familiar with the designs.

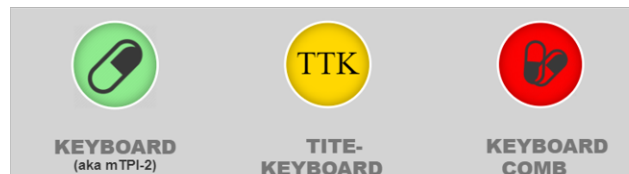
The strengths of the BOIN suite and Keyboard suite lie in the versatile platform and use-friendly features as below.

- 1) Contain multiple novel designs, each addressing a different challenge in Phase I clinical trials.
- 2) Run under different operating systems.
- 3) Do not require any installation.

A. BOIN suite



B. Keyboard suite



C. Choice of design based on trial setting

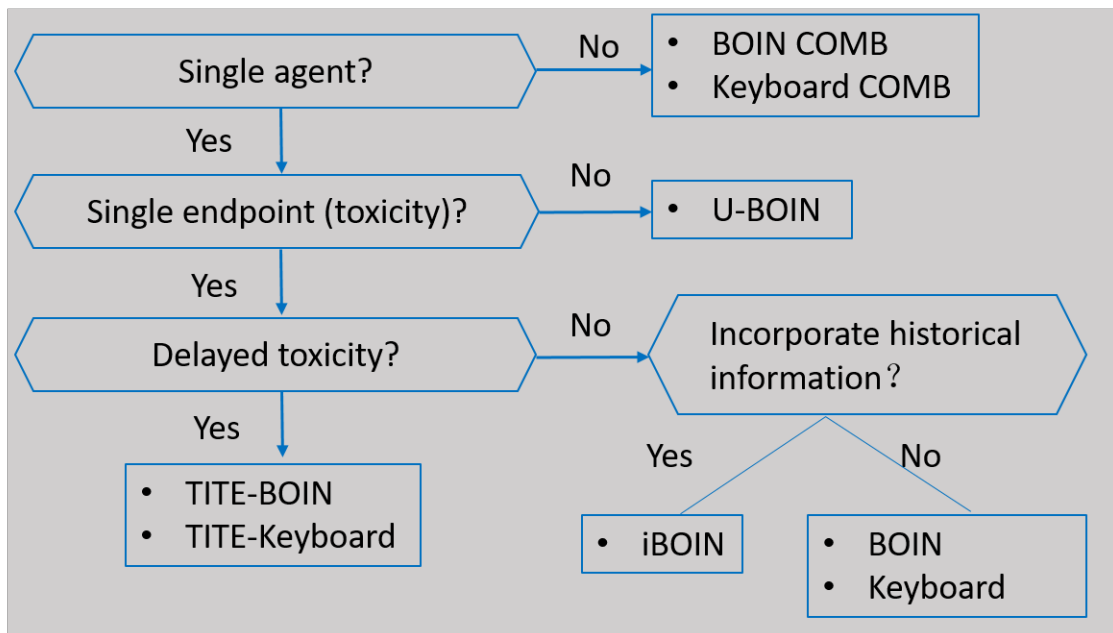


Figure 5.3: Bayesian optimal interval (BOIN) and Keyboard suite available at www.trialdesign.org.

- 4) Do not require statistical programming skills to use.
- 5) Provide comprehensive user help files.
- 6) Regularly maintained and free to use.

All the shiny applications in the two suites have capability to implement the following three main tasks.

Trial planning At the planning stage, users can use the app to conduct simulation for various trial settings and obtain decision rules based on their specified trial setting. All the information is well-organized in a trial protocol template that are available in both html and word format. For the BOIN suite, the templates are available in both English and Chinese.

Trial conduct To determine dose assignment during a trial, users can simply comparing the observed data to the decision boundaries obtained using the apps. Alternatively, for more complicated designs such as U-BOIN, BOIN combination, and Keyboard combination, the apps also provide an option to determine the dose for next cohort of patients. Users can simply type in or upload trial data to the app to get the recommended dose for next cohort of patients.

MTD (or OBD) estimation After a trial completes accrual and has outcomes evaluated, the apps can be used to determine the MTD or OBD.

5.3.2 Implementation with a trial example

All the shiny applications include the following core statistical specifications for trial planning and conduct.

- 1) Target toxicity (or DLT) probability and toxicity intervals required by a design.
- 2) Number of doses and starting dose level.
- 3) Sample size and cohort size.
- 4) Early stopping criteria.
- 5) Prior specifications (applicable for iBOIN, TITE-BOIN, and TITE-KEYBOARD).

To demonstrate the ease of trial planning and trial conduct using the shiny apps, we use the BOIN app (Figure 5.4 shows part of the user interface) as an example. Suppose the target DLT probability in the trial we plan is 0.3. A dose with DLT probability of 0.18 is considered sub-therapeutic and of 0.48 is over-toxic. The information happens to be the default setting of BOIN design, and is used to obtain dose escalation and de-escalation boundaries. In the BOIN app, check the box for “Use the default alternative to minimize decision error (recommended)” to obtain dose escalation and de-escalation boundaries. A maximum of 30 patients will be enrolled and are treated in a cohort size of three. We enter this information in the “Sample Size” panel. To save sample size, we determine that the trial will be stopped if any of the doses have been used

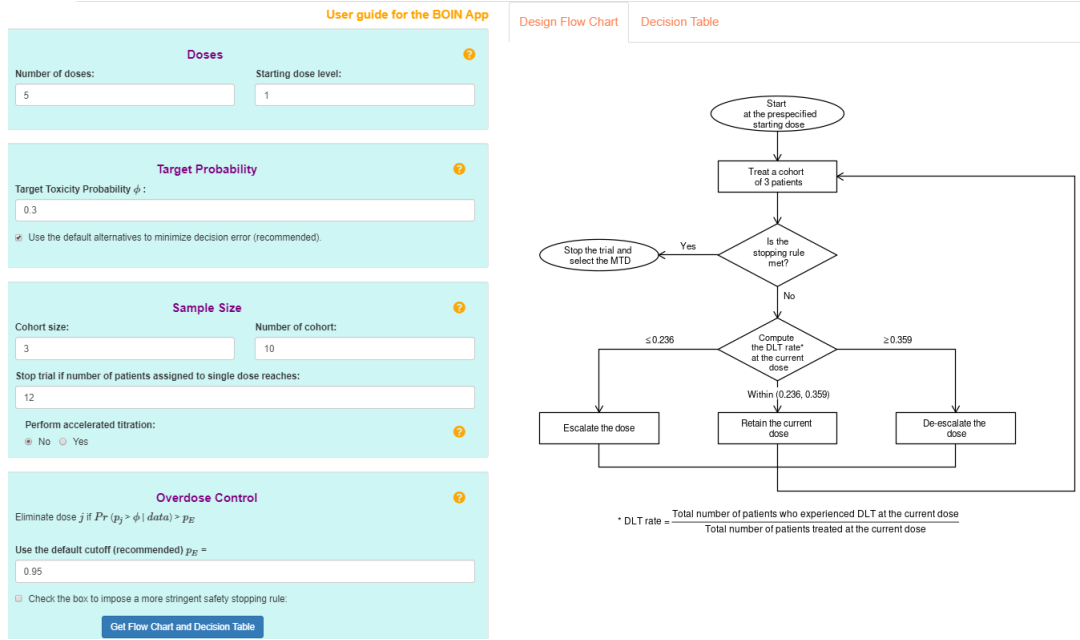


Figure 5.4: Design setting and flowchart for the trial example using the BOIN app.

to treat 12 patients. To prevent from exposing patients to overly toxic doses, we impose the safety rule that a dose and its higher dose levels will be eliminated if we find that there is 95% chance that the dose has a DLT probability that is greater the target probability of 0.3, given at least three patients are treated. The safety criteria is entered under the “Overdose Control” panel.

Trial planning Under the “Trial setting” of the BOIN app, we entered our trial settings as shown in Figure 5.4. Click on “Get Flow Chart and Decision Table”, the app returns the BOIN diagram and decision tables (shown in Figure 5.4 and Figure 5.5, respectively), which can be directly included in trial protocol. ‘

It is critical to assess the accuracy and the safety of design before using it in the actual trial. To do this, we first construct several scenarios that may represent the true dose-toxicity curve. Then we use the “Simulation” tab in the BOIN app to evaluate the probability of correctly selecting the true MTD and the percentage (or number) of patients allocated to each dose. A design with good operating characteristics should have a desirable probability to identify the true MTD and treat a large proportion (or number) of patients at the MTD across various scenarios.

In this example, we consider four scenarios with the MTD located at dose level 1, 2, 3, and 4, respectively, as shown in Figure 5.6. To enter the scenarios, the BOIN app allows a direct input (i.e., type in) and also provides an option for users to upload the scenarios using a prespecified template. If “Upload scenario file” is selected, users will see the template. We choose “Type in” to

Table 1: Dose escalation/de-escalation rule.

	1	2	3	4	5	6	7	8	9	10	11	12
Number of patients treated	1	2	3	4	5	6	7	8	9	10	11	12
Escalate if # of DLT \leq	0	0	0	0	1	1	1	1	2	2	2	2
Deescalate if # of DLT \geq	1	1	2	2	2	3	3	3	4	4	4	5
Eliminate if # of DLT \geq	NA	NA	3	3	4	4	5	5	5	6	6	7

Figure 5.5: Decision table for the trial example using the BOIN app.

enter the scenarios manually. We can conveniently add or remove a scenario through the “Add a Scenario” and “Remove a Scenario” tabs. We can also save our scenarios for future use by clicking on the “Save Scenarios” tab. After scenarios are entered, we set the number of simulated trials (e.g., 1000) for each scenario. To ensure reproducible results of our simulation, we set the random number generator seed as 6 (any other number is fine) under “Set Seed”. At the end, we click “Run Simulation”. On the right panel of the app are displayed the simulation results, including the percentage of selecting a dose as MTD (Selection %) and the percentage of patients treated at each dose (% Pts treated). Additionally, the results also contain the average number of patients (Number of Patients) and the percentage of stopping the trial early without selecting MTD (% Early Stopping). As shown, if we want, we can copy or print the simulation results, or save them in CSV and Excel format. We find this to be a friendly feature that enables easy sharing of the simulation results within a collaboration team. The BOIN app provides a very handy tab to help

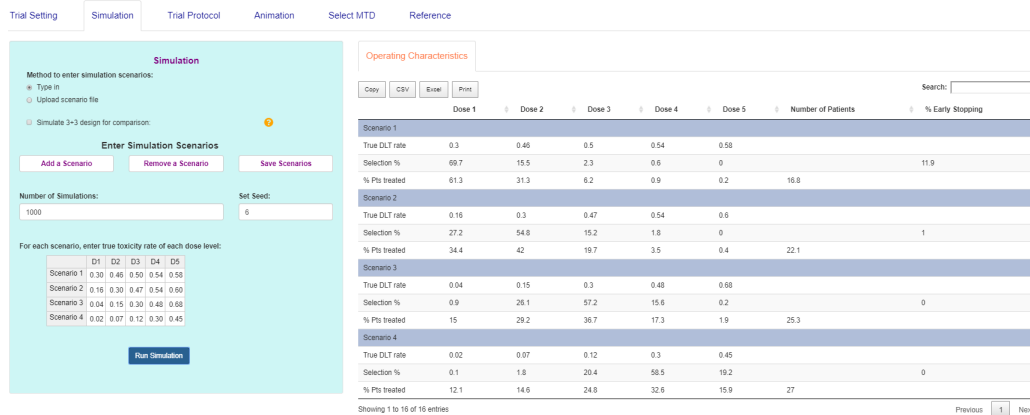


Figure 5.6: Simulation example using the BOIN app.

prepare trial protocol. After running our simulation, we click on “Trial protocol” to download a template to prepare our protocol. The template includes a well-written description of the dose finding algorithm of BOIN design, along with the design diagram, decision table, and simulation results. The template is available in html and word formats. The app also provide Chinese template to meet the ever increasing demand for the BOIN design in China. The English protocol template for our trial example is provided in the Appendix.

Trial conduct To conduct the trial, the decision table (shown in Figure 5.5) obtained during our trial planning is all we need. We treat the first cohort of patients at the lowest dose as specified and use the decision table to determine the dose for subsequent cohorts of patients. Suppose there are no toxicity observed for the first three patients, we will escalate to dose level 2 for next cohort of patients. At any time, we simply count the number of patients treated on that dose and the number of patients who experience toxicity to determine whether escalation or de-escalation is triggered or the current dose should be eliminated. For example, suppose 6 patients were treated at current dose, we would de-escalate the dose if 3 of them experienced toxicity or eliminate the dose and its higher doses if 4 or more patients experienced toxicity. The trial continued until the maximum sample size is reached.

MTD estimation BOIN design uses the isotonic estimate of toxicity rate to select MTD. This can be easily done using the shiny app. Suppose at the end of the trial, we have treated 3, 6, 12, 3 on dose level 1, 2, 3, and 4, but no patients have been treated on dose level 5. The corresponding numbers of patients who experienced toxicity are 0, 1, 3, and 2, respectively. We entered the trial data into the BOIN app under the “Select MTD” tab as shown in Figure 5.7. A click on “Estimate the MTD” will give us the selection result, as shown on the right side of the app, where we see the estimated MTD, and the dose-limiting toxicity (DLT) estimate for all doses used to treat patients and their corresponding 95% credible intervals.

5.4 Summary

To facilitate the use of the novel model-assisted designs and accelerate drug development in early phase trials, we developed two versatile software platforms: BOIN suite and Keyboard suite. Each suite includes multiple well-performed model-assisted designs that can be used to tackle different challenges in Phase I clinical trials. The availability of these software provides timely and interactive simulation information to help clinicians and reviewers to understand the

accuracy and the safety of the novel designs without the need of programming skills. This work is of greatest importance, as it not only facilitates the ease of collaborations between statisticians and clinicians, but also will advance the development of novel methods in clinical trials by making them more accessible to a broader audience.

A Appendix

The protocol template (a pdf file) for the trial example using BOIN app is attached below. This file demonstrates how the app can save the time used to write a statistical protocol by using the readily available template from the shiny app.

NOTE: If you would like a Microsoft Word version of this protocol document template, simply select (highlight) this entire document, copy it, open a new Word document, and paste the protocol document template into the Word document. On Windows machines this can conveniently be accomplished by pressing CTRL-A to select (highlight) the entire document, CTRL-C to copy it, and CTRL-V to paste it. Alternatively, opening the saved HTML file in Microsoft Word may produce better results.

Template for Protocol Preparation

We will employ the Bayesian optimal interval (BOIN) design (Liu and Yuan, 2015; Yuan et al., 2016) to find the MTD. The BOIN design is implemented in a simple way similar to the traditional 3+3 design, but is more flexible and possesses superior operating characteristics that are comparable to those of the more complex model-based designs, such as the continual reassessment method (CRM) (Zhou et al., 2018).

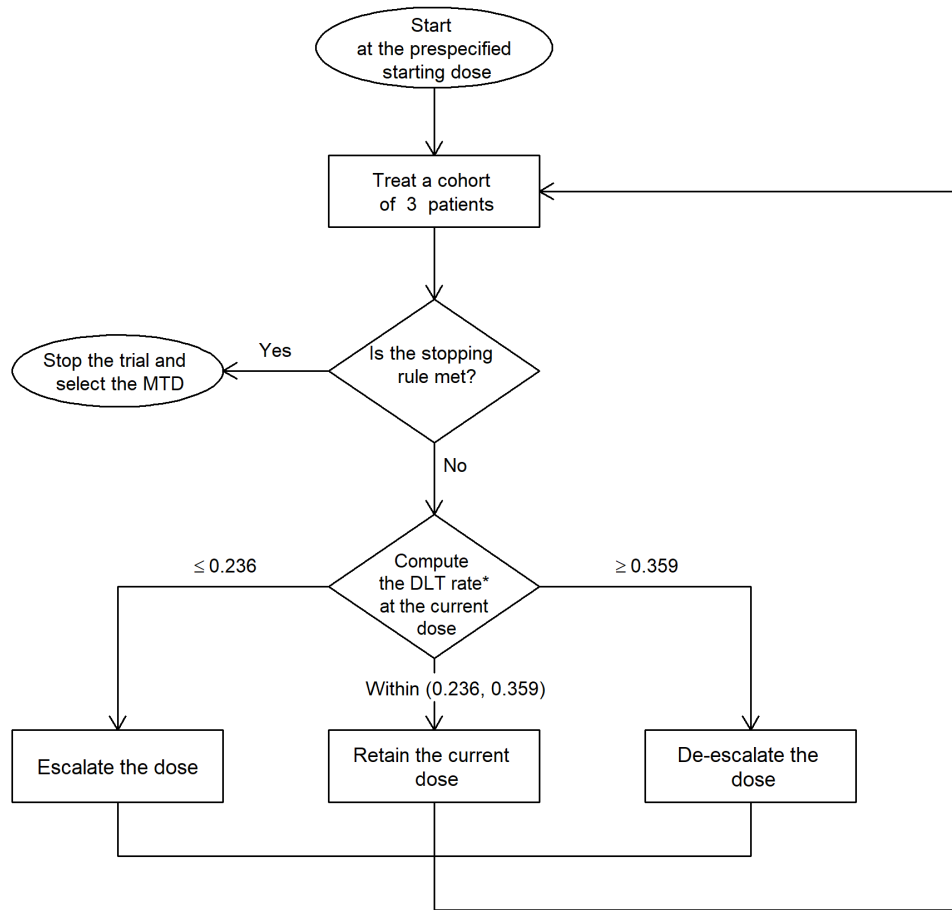
The target toxicity rate for the MTD is $\phi = 0.3$ and the maximum sample size is 30. We will enroll and treat patients in cohorts of size 3. DLTs are defined in **Section ###**, and only those DLTs that occur within **the first cycle** will be used for dose finding. As shown in Figure 1, the BOIN design uses the following rule, optimized to minimize the probability of incorrect dose assignment, to guide dose escalation/de-escalation:

- if the observed DLT rate at the current dose is ≤ 0.236 , escalate the dose to the next higher dose level;
- if the observed DLT rate at the current dose is ≥ 0.359 , de-escalate the dose to the next lower dose level;
- otherwise, stay at the current dose.

For the purpose of overdose control, doses j and higher levels will be eliminated from further examination if $\Pr(p_j > 0.3 \mid \text{data}) > 0.95$, where p_j is the true DLT rate of dose level j , $j = 1, \dots, 5$. This posterior probability is evaluated based on the beta-binomial model $y_j \mid p_j \sim \text{binomial}(p_j)$ with $p_j \sim \text{uniform}(0, 1)$, where y_j is the number of patients experienced DLT at dose level j . When the lowest dose is eliminated, stop the trial for safety. The probability cutoff 0.95 is chosen to be consistent with the common practice that when the target DLT rate $\leq 1/6$, a dose with 2/3 patients experienced DLT is eliminated.

The steps to implement the BOIN design are described as follows:

1. Patients in the first cohort are treated at dose level 1.
2. To assign a dose to the next cohort of patients, conduct dose escalation/de-escalation according to the rule displayed in Table 1, which is equivalent to the rule described above. When using Table 1, please note the following:
 - a. "Eliminate" means eliminate the current and higher doses from the trial to prevent treating any future patients at these doses because they are overly toxic.
 - b. When we eliminate a dose, automatically de-escalate the dose to the next lower level. When the lowest dose is eliminated, stop the trial for safety. In this case, no dose should be selected as the MTD.
 - c. If none of the actions (i.e., escalation, de-escalation or elimination) is triggered, treat the new patients at the current dose.
 - d. If the current dose is the lowest dose and the rule indicates dose de-escalation, treat the new patients at the lowest dose unless the number of DLTs reaches the elimination boundary, at which point terminate the trial for safety.
 - e. If the current dose is the highest dose and the rule indicates dose escalation, treat the new patients at the highest dose.
3. Repeat step 2 until the maximum sample size of 30 is reached, or stop the trial if the number of patients treated at the current dose reaches 12 and the decision according to Table 1 is to stay at the current dose.



$$* \text{ DLT rate} = \frac{\text{Total number of patients who experienced DLT at the current dose}}{\text{Total number of patients treated at the current dose}}$$

Figure 1. Flowchart for trial conduct using the BOIN design

Table 1. Dose escalation/deescalation rule for the BOIN design

	The number of patients treated at the current dose											
	1	2	3	4	5	6	7	8	9	10	11	12
Escalate if # of DLT ≤	0	0	0	0	1	1	1	1	2	2	2	2
Deescalate if # of DLT ≥	1	1	2	2	2	3	3	3	4	4	4	5
Eliminate if # of DLT ≥	NA	NA	3	3	4	4	5	5	5	6	6	7

Note. # of DLT is the number of patients with at least 1 DLT. When none of the actions (i.e., escalate, de-escalate or eliminate) is triggered, stay at the current dose for treating the next cohort of patients.

After the trial is completed, select the MTD based on isotonic regression as specified in Liu and Yuan (2015). This computation is implemented by the shiny app "BOIN" available at <http://www.trialdesign.org> (<http://www.trialdesign.org>). Specifically, select as the MTD the dose for which the isotonic estimate of the toxicity rate is closest to the target toxicity rate. If there are ties, select the higher dose level when the isotonic estimate is lower than the target toxicity rate and select the lower dose level when the isotonic estimate is greater than or equal to the target toxicity rate.

Operation Characteristics

Table 2 shows the operating characteristics of the trial design based on 1000 simulations of the trial using shiny app "BOIN" available at <http://www.trialdesign.org> (<http://www.trialdesign.org>). The operating characteristics show that the design selects the true MTD, if any, with high probability and allocates more patients to the dose levels with the DLT rate closest to the target of 0.3.

Table 2. Operating characteristics of the BOIN design

	Dose Level					Number of Patients	% Early Stopping
	1	2	3	4	5		
<u>Scenario 1</u>							
True DLT Rate	0.3	0.46	0.5	0.54	0.58		
Selection %	69.7	15.5	2.3	0.6	0		11.9
% Pts Treated	61.3	31.3	6.2	0.9	0.2	16.8	
<u>Scenario 2</u>							
True DLT Rate	0.16	0.3	0.47	0.54	0.6		
Selection %	27.2	54.8	15.2	1.8	0		1
% Pts Treated	34.4	42	19.7	3.5	0.4	22.1	
<u>Scenario 3</u>							
True DLT Rate	0.04	0.15	0.3	0.48	0.68		
Selection %	0.9	26.1	57.2	15.6	0.2		0
% Pts Treated	15	29.2	36.7	17.3	1.9	25.3	
<u>Scenario 4</u>							
True DLT Rate	0.02	0.07	0.12	0.3	0.45		
Selection %	0.1	1.8	20.4	58.5	19.2		0
% Pts Treated	12.1	14.6	24.8	32.6	15.9	27	

Reference

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

Yuan Y., Hess K.R., Hilsenbeck S.G. and Gilbert M.R. (2016) Bayesian Optimal Interval Design: A Simple and Well-performing Design for Phase I Oncology Trials, *Clinical Cancer Research*, 22(17), 4291-4301.

Zhou, H., Yuan, Y., & Nie, L. (2018). Accuracy, safety, and reliability of novel phase I trial designs. *Clinical Cancer Research*, 24(18), 4357-4364.

Appendix 1: Trial and Design Specifications

Parameter	Value
Number of doses	5
Starting dose	1
Max sample size	30
Cohort size	3
Stop trial if # patients assigned to single dose reaches	12
Use accelerated titration	False
Target toxicity probability	0.3
Use the default alternatives to minimize decision errors	True
Alternative (unacceptable high toxicity) for optimization	Default
Alternative (unacceptable low toxicity) for optimization	Default
Eliminate dose threshold	0.95
Number of repetitions per scenario	1000
Random number generator seed	6

Appendix 2. Alternative BOIN decision table

		Number of patients treated at current dose											
		1	2	3	4	5	6	7	8	9	10	11	12
No. of patients with DLT	0	E	E	E	E	E	E	E	E	E	E	E	E
	1	D	D	S	S	E	E	E	E	E	E	E	E
	2		D	D	D	D	S	S	S	E	E	E	E
	3			DE	DE	D	D	D	D	S	S	S	S
	4				DE	DE	DE	D	D	D	D	D	S
	5					DE	DE	DE	DE	DE	D	D	D
	6						DE	DE	DE	DE	DE	DE	D
	7							DE	DE	DE	DE	DE	DE
	8								DE	DE	DE	DE	DE
	9									DE	DE	DE	DE
	10										DE	DE	DE
	11											DE	DE
	12												DE

■ E=Escalate to the next higher dose
■ S=Stay at the current dose
■ D=De-escalate to the next lower dose
■ DE=De-escalate and eliminate the current and higher doses

Target Toxicity Probability ϕ :

0.3

Number of doses

5

Please enter the trial data:

Dose level	Number of patients treated	Number of patients with dose limiting toxicity
1	3	0
2	6	1
3	12	3
4	3	2
5	0	0

Estimate the MTD

MTD Selection Result

The MTD is dose level 3

Dose Level	Posterior DLT Estimate	95% Credible Interval	Pr(toxicity>0.3 data)
1	0.02	(0.00 , 0.20)	0.01
2	0.17	(0.01 , 0.53)	0.18
3	0.25	(0.06 , 0.52)	0.32
4	0.66	(0.16 , 0.99)	0.91
5	----	(-----)	----

NOTE: no estimate is provided for the doses at which no patient was treated.

Figure 5.7: Design setting of trial example and flowchart produced in the BOIN app.

CHAPTER 6

The use of local and nonlocal priors in Bayesian test-Based monitoring for Phase II clinical trials

6.1 Introduction

Phase II clinical trials play a critical role in drug development. The objective of phase II clinical trials is to evaluate the therapeutic effect of a new treatment and screen out inefficacious agents. Phase II trials commonly use a short-term and dichotomous endpoint to characterize the patient clinical response to treatment. For example, a binary objective response endpoint can be used to indicate whether the patients have achieved a complete or partial response within a predefined treatment course. If the new treatment shows a promising treatment effect, further large-scale confirmatory trials can be conducted.

In general, phase II studies can be performed using either frequentist or Bayesian designs [4, 117]. The frequentist approaches treat the response probability of the new treatment as a fixed, yet unknown parameter [9, 23, 26]. The most well-known frequentist phase II design is the Simon's optimal two-stage design [88]. In this design, n_1 patients are enrolled in stage I, if there are r_1 or fewer responses, terminate the trial early and conclude that the drug is not promising; otherwise, enroll an additional n_2 patients in stage II and claim the drug is promising if more than r responses are observed. The parameters of Simon's design are optimized to minimize the expected sample size (the optimal design) or the maximum sample size (the minimax design) under the null hypothesis that the treatment is not efficacious, with the type I and type II error rates being well controlled. Simon's two-stage design is widely adopted in phase II trials. Numerous generalizations or extensions of Simon's two stage design have been proposed. Examples include the three-stage design [11], the design based on a composite null hypothesis [59], the optimal adaptive design [87],

and the multiple-arm design [123] among others.

Most of the frequentist phase II designs only have two or three stages, with the design parameters determined based on numerical searching. However, when the number of interim looks exceeds three or continuous monitoring is considered, it is computationally infeasible to determine the optimal design parameters for frequentist designs. Moreover, frequentist designs are rigid in the sense that they do not allow any deviation from the original design. For example, for the Simon’s design, the “go/no-go” decisions can only be made at the predetermined interim time, and the design will not stop the trial early, even when observed data have already indicated that the drug is indeed ineffective. To safe guard patients from futile treatments and accelerate the development of efficacious drugs, it is critical to make adaptive decisions based on the accumulated data throughout the trial in a timely fashion, such that the trial can be stopped earlier for futility or efficacy. Typically, the more interim looks, the greater the chance of detecting futile or efficacious treatments; in turn, the smaller the sample size, the more efficient the trial.

When more interim analyses are desired, Bayesian methods become particularly appealing due to their “we learn as we go” nature [39]. Bayesian inference universally follows three steps: (1) Elicit the prior distribution of the unknown response parameter, based on historical or external trials, as well as expert opinions. (2) Obtain the likelihood function based on the data collected during the study. (3) Synthesize the prior information and the observed data likelihood into a posterior distribution of the parameter of interest using Bayes theorem. In Bayesian phase II designs, the treatment effect is considered as an unknown random variable, and the Bayesian inference can be adaptively and timely made, as long as new data are observed. As a result, Bayesian designs usually are more flexible in facilitating multiple interim decisions. In addition, frequentist methods require a penalty for each look of the data, but there are no such hurdles for interim analyses in a Bayesian perspective [3].

Due to the appealing feature of flexibility, recent years have witnessed vast developments in Bayesian phase II clinical trial designs [32, 45, 49, 86, 94, 98, 99, 103, 107, 122]. Overall, existing Bayesian sequential monitoring approaches can be divided into two categories: posterior (or predictive) probability based (PB) or Bayesian hypothesis test based (TB). For the PB approach, the “go/no-go” decisions are made based on the Bayesian posterior (or predictive) probability that the treatment is futile or effective. If such a probability is greater than a prespecified upper probability cutoff, then the trial can be terminated early, as the experimental treatment is likely to be futile or promising. If this probability is small, however, then there is inadequate information to deliver

a conclusion and the trial continues to collect more data. On the other hand, the TB approach adaptively makes the “go/no-go” decisions based on the Bayesian hypothesis testing framework with the Bayes factor. Although [45] argued that the use of Bayes factor in sequential monitoring can gain more efficiency compared to the PB approach, while also eliminating a potential source of bias often caused by prior-data conflict, the research on the TB approach for phase II trials is lacking. Limited examples include the two-stage Bayes factor based design [20] and the design to identify the maximum effective dose [30]. To get more insight into TB monitoring, in the first part of this chapter, we are interested in studying the connection between the TB and PB approaches. In particular, we show that TB sequential monitoring is essentially a special case of PB monitoring, with a data-adaptive stopping cutoff.

Although the Bayesian approaches have gained popularity in phase II clinical trials, an issue that was largely overlooked is the choice of prior distributions. For the TB monitoring approach, a prior distribution is required to quantify the initial uncertainty of the unknown parameter under the alternative hypothesis. According to [46], the prior distributions can be classified into local priors and nonlocal priors. A local prior refers to a probability density that assigns positive probability to regions of the parameter space that are consistent with the null hypothesis. The commonly used Beta distribution for the response rate falls into the category of local priors. [46] also showed that the problem with using the local prior is that it accumulates evidence in favor of the true null hypothesis at a much slower rate, as compared to favoring a true alternative hypothesis. To address this issue, a class of nonlocal prior was proposed. One such nonlocal prior is the flexible inverse moment (iMOM) density [45, 46]. In the second part of this paper, we extensively examine the effects of local and nonlocal priors on the TB monitoring procedure. We demonstrate that when the treatment is truly futile, using an iMOM prior has a greater probability of stopping a trial early for futility vs using a Beta prior, while the probabilities to claim efficacy when the treatment is truly efficacious are comparable between the two. This, in turn, leads to a smaller average number of patients enrolled when the iMOM prior is used.

The rest of this chapter is organized as follows. In section 6.2, we first provide an overview of different Bayesian monitoring approaches and then show the relationship between PB monitoring and TB monitoring. In section 6.3, we describe prior specification for Beta distribution and iMOM density, and we propose a procedure to estimate the associated effective sample size for iMOM density. In section 6.4, we conducted extensive simulation to examine the operating characteristics of the TB monitoring using Beta and iMOM priors. In section 6.5, we provide tutorials on the

newly developed web-based software for phase II sequential monitoring. We conclude the study in section 6.6.

6.2 Sequential monitoring

6.2.1 Bayesian posterior (or predictive) probability based (PB) monitoring

In a phase II trial, let n denote the number of patients and y denote the number of responses observed. The treatment effect is quantified by the probability of clinical response, p , and thus the number of response y is usually modeled as a binomial random variable, i.e., $y \sim \text{Binom}(n, p)$. Given the observed value $D = (y, n)$, the binomial likelihood function is $f(D | p) = \binom{n}{y} p^y (1 - p)^{n-y}$. In Bayesian inference, all the information about the parameter is contained in the posterior distribution of p , which can be calculated using the Bayes' theorem as follows,

$$(6.1) \quad \pi(p | D) = \frac{f(D | p)\pi(p)}{\int f(D | p)\pi(p)dp}, \quad 0 \leq p \leq 1.$$

Here, $\pi(p)$ is the prior distribution for the unknown parameter p , which characterizes all available information before conducting the trial. Once the posterior distribution is obtained, we can make inference about the true parameter p .

Let p_0 denote a clinically uninteresting response rate or the response rate of the standard control, and let p_1 represent a desirable target response rate. In general, the “go/no-go” decisions for the PB approach can be made based on either the posterior probability or the predictive probability that the experimental drug has a higher response rate than the clinically uninteresting rate p_0 . Specifically, if posterior probability is used, the posterior probability that the treatment effect of the experimental drug exceeds the standard effect by a prespecified improvement δ , i.e., $\Pr(p > (p_0 + \delta) | D)$, can be calculated continuously at each interim decision-making time when the data are accumulated. Here, we can assign either a point mass or a prior probability distribution based on historical data to p_0 . In this paper, we only consider p_0 as a prespecified point, and fix $\delta = 0$.

For a trial with sequential monitoring, we can determine whether to early stop or continue the trial by comparing the posterior probability $\Pr(p < (p_0 + \delta) | D)$ with a probability cutoff [98]. For example, the trial can be stopped early for futility if $\Pr(p < (p_0 + \delta) | D) > r_f$, where r_f is a probability cutoff for futility stopping. If this Bayesian monitoring rule is not satisfied, then there is not adequate information to deliver any conclusion, so the trial continues to collect more data until the prespecified sample size is reached. Similarly, if efficacy monitoring is considered, then the

trial can be stopped early for efficacy, if the posterior probability $\Pr(p > (p_0 + \delta) \mid D)$ is greater than the efficacy stopping boundary r_e . In most cases, the probability cutoffs r_f and r_e take fixed values at 0.90 or 0.95. Alternatively, calibration procedures can also be conducted to calibrate the cutoff values of r_f and r_e to satisfy certain prespecified type I/II error constraints. Generalization of this Bayesian monitoring procedure can be found in [32], [99], and [122] among others.

In addition to the posterior probability, Bayesian predictive probability serves as a good alternative tool to conduct interim decision monitoring. Unlike the posterior probability monitoring approach, which is based on the posterior distribution of p , the predictive probability monitoring procedure uses the predictive distribution of p for decision making. Suppose the maximum sample size is N , at the interim time with n patients treated, let Y^* denote the future number of responses among the remaining $N - n$ patients. Given the observed data, the predictive distribution of the number of responses among the total sample size can be obtained as $\Pr(Y^* \mid D, N - n) = \int_0^1 f(D^* \mid p)\pi(p \mid D)dp$, where $D^* = (Y^*, N - n)$ denotes the unobserved future data, $f(D^* \mid p)$ is the Binomial likelihood function based on the future data D^* , and $\pi(p \mid D)$ is the posterior distribution of p based on the observed data D .

The Bayesian predictive probability procedure continuously calculates the predictive probability of a positive conclusion (e.g., accept the alternative hypothesis) or of a negative conclusion (e.g., accept the null hypothesis), and makes timely adaptive “go/no-go” decisions as the trial moves on [49]. By continuously comparing the predictive probabilities with respect to the cutoffs, the decision rules of the predictive probability monitoring procedure mimics that of the posterior probability monitoring procedure. As shown in [49], the sequential monitoring based on the predictive probability not only accounts for uncertainty among observed data, but also variability for outcomes that may be observed in the future for a trial. Like posterior probability monitoring, the predictive probability monitoring is more adaptable than traditional multi-stage designs and conceptually appealing. It has been generated to more complex trial settings, such as randomized phase II trials [107], trials with time-to-event endpoints [108], and platform trials [34].

6.2.2 Bayesian hypothesis test based (TB) monitoring

An alternative Bayesian monitoring approach is developed on the basis of the Bayesian hypothesis testing procedure [45]. Specifically, consider the following hypotheses,

$$(6.2) \quad H_0 : p \leq p_0 \quad \text{versus} \quad H_1 : p > p_0,$$

where under the null hypothesis H_0 , the treatment effect is deemed unpromising. Under the Bayesian framework, the Bayes factor serves as a natural quantity to measure the strength of evidence for a hypothesis relative to another one. Specifically, the Bayes factor in favor of H_1 (denoted as BF_{10}) can be calculated as

$$(6.3) \quad BF_{10} = \frac{\Pr(D \mid H_1)}{\Pr(D \mid H_0)} = \frac{\int_{p_0}^1 f(D \mid p)\pi(p \mid H_1)dp}{\int_0^{p_0} f(D \mid p)\pi(p \mid H_0)dp},$$

where $f(D \mid p)$ is the data likelihood and $\pi(p \mid H_j)$ is the prior distribution of p under H_j , $j = 0, 1$. That is, the Bayes factor is the ratio of likelihood averaged over the prior under H_1 and H_0 . When no or little prior information can be borrowed, it is natural to assume *a priori* that the hypotheses are equally likely *a priori*, i.e., the prior hypothesis probabilities $\Pr(H_0) = \Pr(H_1) = 1/2$. In this case, the Bayes factor BF_{10} is equivalent to the posterior odds in favor of H_1 .

Under the Bayesian hypothesis testing framework, a decision to reject the null hypothesis occurs only when the Bayes factor in favor of the alternative hypothesis exceeds a prespecified evidence level. Thus, it can quantify evidence in favor of the null hypothesis when the null hypothesis is not rejected [55]. A greater BF_{10} indicates a stronger evidence for H_1 . Alternatively, we can define Bayes factor as $BF_{01} = \Pr(D \mid H_0)/\Pr(D \mid H_1)$. In this case, a larger BF_{01} indicates a greater evidence for H_0 .

At each interim decision-making time, the TB monitoring procedure adaptively tests the hypotheses (6.2) and makes the “go/no-go” decisions as follows:

- Efficacy stopping: if $BF_{10} > \gamma_e$, stop the trial to claim efficacy of the new treatment.
- Futility stopping: if $BF_{01} > \gamma_f$, stop the trial to claim futility of the new treatment.

where γ_e and γ_f are prespecified values that respectively represent substantial evidence in favor of H_1 and H_0 . Table A6.1 shows the interpretation of Bayes factor as the strength of evidence for hypothesis testing [47]. [45] showed that the Bayes factor in favor of the alternative hypothesis will almost always be smaller than it could have been with a correctly-specified prior under H_1 . As a result, compared to the PB monitoring that is based on the posterior probability intervals, the TB monitoring based on Bayes factor can automatically adjust the bias induced by overly optimistic priors. This property is of practical use, especially from a regulatory perspective.

6.2.3 Relationship between PB monitoring and TB monitoring

As aforementioned, PB monitoring is based on the posterior probability that the response rate is greater than the reference response rate given observed data, while the TB monitoring is based on the evidence contained in the data in favor of one hypothesis relative to another one. We show here that the two monitoring approaches are closely related, and TB monitoring can essentially be viewed as a special case of PB monitoring. Let $\pi(p) = \Pr(H_0)\pi(p | H_0) + \Pr(H_1)\pi(p | H_1)$ be the overall prior distribution, and let $\pi(p | D)$ be the overall posterior distribution of p , which is calculated based on the observed data D and the overall prior $\pi(p)$. Then the posterior probability of H_j can be computed as $\Pr(H_j | D) = \int_{H_j} \pi(p | D)dp$, $j = 0, 1$. According to the definition of Bayes factor, for futility monitoring, we have

$$\begin{aligned}
& \frac{\Pr(D | H_0)}{\Pr(D | H_1)} > \gamma_f \\
\iff & \frac{\Pr(H_0 | D) \Pr(H_1)}{\Pr(H_1 | D) \Pr(H_0)} > \gamma_f \\
\iff & \Pr(H_0 | D) > \frac{\gamma_f \Pr(H_0) \Pr(H_1 | D)}{\Pr(H_1)} \\
\iff & \Pr(p \leq p_0 | D) > \frac{\gamma_f \Pr(H_0) \Pr(p > p_0 | D)}{\Pr(H_1)} \equiv r'_f(D)
\end{aligned}$$

Here, $r'_f(D)$ depends on the observed data D . Thus TB monitoring can be viewed as a special case of PB monitoring with $\delta = 0$ and a data-adaptive stopping cutoff. A similar proof can be done for efficacy monitoring.

Compared to the fixed cutoff that remains unchanged over the course of the trial, the data-adaptive cutoff $r'_f(D)$ varies with the interim sample size n and the observed number of responses y . For example, suppose the data are generated under the null hypothesis H_0 , then in general, the cutoff function $r'_f(D)$ decreases with the number of patients treated. This is due to the fact that the posterior probability $\Pr(p > p_0 | D)$ decreases as the sample size increases when the data are simulated from H_0 . As a result, at the beginning of the trial, data are sparse and a more stringent stopping rule with a larger value of $\lambda'_f(D)$ may be preferred to avoid accidentally terminating the trial. As the trial proceeds and information accumulates, there is less uncertainty regarding the response rate and a smaller stopping boundary would be produced based on $\Pr(p > p_0 | D)$, making the trial easier to stop for futility. This is similar to Bayesian posterior monitoring, in which the probability cutoff λ_e is generalized into decreasing function of sample size [98]. [122] showed that allowing the cutoff to change with the sample size increases the power of a design. Indeed, [45]

showed an extensive simulation study where TB monitoring exhibited higher efficiency than PB monitoring.

6.3 Prior specification

Prior specification is critical to the performance of Bayesian inference. Especially in early-phase trials when the sample size is typically small, different prior distributions may lead to varied interim decisions or final conclusions [110]. During a study, the Bayesian method updates the prior to form the posterior distribution coherently and continuously based on the accumulated data. The use of prior information for Bayesian designs could be a double-edged sword. If the prior is consistent with the current data, proper use would increase the trial efficiency and render more accurate decision making. In contrast, improper use of the prior may be subject to biased inference and incorrect decisions. Hence, the choice of prior distribution should be carefully made according to the clinical setting, expert knowledge, and sufficient simulation studies. Recently, the US Food and Drug Administration [22] has also announced a draft guideline for industry and highlighted the importance of the evaluation of prior distribution on innovative trial designs. For Bayesian phase II trial designs, a largely overlooked issue is the choice of prior distributions. In this section, we describe two different types of prior distributions that will be thoroughly examined in the TB monitoring procedure.

6.3.1 Beta distribution

Since the primary endpoint in phase II trials is usually binary, a natural choice of the prior for the response rate (p) is the Beta prior due to the conjugacy. The use of a conjugate prior simplifies the computation for posterior distribution, as it gives a closed-form expression for the posterior. Suppose that the prior distribution is $p \sim \text{Beta}(a, b)$, where a and b are the hyperparameters. The quantity $a/(a + b)$ reflects the prior mean, while the size of $a + b$ indicates how informative the prior is. The larger the value of $a + b$, the more informative the prior, indicating a stronger belief. In many cases, a $\text{Beta}(1, 1)$ is used as a non-informative prior, which is equivalent to the Uniform (0,1) distribution. Given the observed data $D = (y, n)$, the posterior distribution of p is still a Beta distribution, i.e., $p \mid D \sim \text{Beta}(y + a, n - y + b)$. Such a beta-binomial model has been widely applied in early-phase trial designs [91, 49, 59, 40, 33].

6.3.2 The inverse moment (iMOM) density

All beta densities assign non-negligible probability to the parameter space that is consistent with the null hypothesis. Thus, the beta distribution is the so-called local prior. In Bayesian hypothesis testing using the Bayes factor, [46] showed that the problem of using the local prior is that it accumulates evidence in favor of the true null hypothesis at a much slower rate vs a true alternative hypothesis. To address this problem, the nonlocal priors that assign zero prior probability to the null space are advocated. One representative example is the class of inverse moment (iMOM) prior densities. In particular, when the parameter of interest is response rate (p), the iMOM density takes the form:

$$(6.4) \quad \pi(p; p_0, k, \nu, \tau) = \frac{k\tau^{\nu/2}}{\Gamma(\nu/2k)} [(p - p_0)^2]^{-(\nu+1)/2} \exp \left\{ - \left[\frac{(p - p_0)^2}{\tau} \right]^{-k} \right\}$$

where p_0 is the null value, $p \in [0, 1]$, and $k > 0, \nu > 0, \tau > 0$ are the parameters that jointly define the iMOM density. The iMOM density is not a conjugate prior, as the posterior distribution does not have a standard distributional form. Inference on the unknown parameter p , if needed, can be made by drawing posterior samples using Markov Chain Monte Carlo (MCMC) algorithms.

The iMOM density assigns zero density at the point null hypothesis and its neighborhood. We refer to the area that has zero mass as the “null region” hereafter. The null region increases with k and τ and decreases with ν (as shown in Figure A6.1). In terms of the tail behavior, a heavier tail is accompanied by a larger value of τ or a smaller value of ν .

One difference between iMOM and Beta is their robustness to Bayesian inference when prior-data conflicts exist. [25] showed that, in Bayesian inference, a Beta prior is not a robust prior in comparison to a Cauchy prior when there is prior-data conflict. We also checked the iMOM against the Beta and Cauchy priors when a prior-data conflict exists, using the same settings in [25], and we found that iMOM has similar robustness to Cauchy, indicating that iMOM is a more robust prior than Beta in the prior-data conflict setting (Figure A6.2).

6.3.3 Prior effective sample size

A fundamental aspect of using the Bayesian approach is to quantify the amount of information contained in the prior [70]. This is measured by effective sample size (ESS), which is defined as the number of hypothetical patients associated with the prior distribution. It is straightforward

to determine ESS in many commonly used models. For example, it can be argued that a $\text{Beta}(a, b)$ distribution has an ESS of $a + b$. While it is straightforward to determine ESS for a Beta distribution, it is not obvious for an iMOM prior. Following [69], we use the following procedure to obtain ESS for an iMOM prior $\pi(p; p_0, k, \tau, \nu)$.

1. Specify the mode (p_m) of the iMOM prior $\pi(p; p_0, k, \tau, \nu)$ or determine it using the following equation:

$$p_m = p_0 + \sqrt{\tau} \left[\frac{2k}{\nu + 1} \right]^{1/2k}.$$

2. Sample T (e.g., $T = 10^4$) data points $(p^{(1)}, \dots, p^{(T)})$ from $\pi(p; p_0, k, \tau, \nu)$ and calculate the variance of the samples

$$\sigma_p^2 = \frac{\sum_{t=1}^T (p^{(t)} - \bar{p})^2}{T - 1}, \text{ where } \bar{p} = \frac{\sum_{t=1}^T p^{(t)}}{T}.$$

3. Specify a $\text{Beta}(c, d)$ such that the mode and variance of this Beta distribution matches those of the iMOM prior, i.e., set

$$p_m = \frac{c - 1}{c + d - 2}, \text{ and } \hat{\sigma}_p^2 = \frac{cd}{(c + d)^2(c + d + 1)}.$$

4. Solve the two nonlinear equations in *step 3* to obtain c and d , and determine effective sample size as

$$\text{ESS} = c + d.$$

One of the main tasks in our study is to assess the influence of different priors on the operating characteristics of a phase II study with TB monitoring based on the Bayes factor. For a fair comparison, we first make sure that the priors have the same ESS, given that they have the same mean or mode. It is intuitive and easy to specify a Beta prior with a desired ESS. For instance, if a Beta prior with mode 0.5 and ESS=10 is needed, a $\text{Beta}(5, 5)$ will satisfy the specification.

The corresponding specification for iMOM, however, is not straightforward. In this study, we use a numerical calibration procedure to specify an iMOM prior. First, we specify the following three elements for the iMOM prior: (1) the uninteresting response rate p_0 under H_0 ; (2) the prior mode (p_m) of response rate for the new treatment under H_1 ; and (3) the prior effective sample size. Based on the three prefixed quantities outlined above, we determine the (k, τ, ν) for the iMOM prior through a grid search. Specifically, we set up a wide range for k (e.g., $k \in (0.01, 10)$), and $\nu = 2k$,

as recommended in [45, 46]. Each combination of k and ν , along with p_0 and p_m , determines a unique τ , and thus a unique iMOM density. Using methods outlined in steps 1-4 above, the values of (k, ν, τ) with $\text{ESS} = c + d$ closest to the prespecified ESS will be selected to determine the iMOM prior.

To facilitate the use of the iMOM in phase II TB monitoring, we develop an easy-to-use web-based application, which is freely available at www.trialdesign.org. This application can be used to conduct interim monitoring for efficacy, futility, or both, based on the Bayes factor. More details on this application are provided in Section 6.5.

6.4 Simulation

6.4.1 Simulation setting

In this paper, we studied the prior influence on the performance of TB monitoring procedure, while the comparisons between TB and PB monitoring have been extensively made in [45] and thus are omitted. On the other hand, our simulation study can be viewed as a complement to the study in [45].

A new drug is under development with the expectation to improve the likelihood of increasing patients' response to a treatment. The response rate for a standard of care was 0.2 and the new treatment was anticipated to have a larger response rate of 0.4, based on external trial information and preliminary results. A trial was planned and the investigator intended to use the TB monitoring design to conduct this trial. The trial was expected to (1) stop early to safeguard patients from futile drug, if the treatment is not promising compared to the standard of care; (2) stop early to accelerate the development of the new treatment, if early evidence indicates highly possibility of success; and (3) enroll fewer patients, given the correct decision is made. To help the investigator design this trial with a suitable prior, we conducted extensive simulations to assess the operating characteristics of 10,000 simulated trials for various scenarios under different prior specifications.

Assume that the trial would enroll at most N patients ($N = 50, 100$). Patients entered the trial with a cohort size of five. The first interim analysis was carried out when 10 patients were treated and had their outcomes evaluated, then an interim analysis was conducted after every five additional patients. The trial was early terminated for efficacy if $BF_{10} \geq 9$; early terminated for futility if $BF_{01} \geq 9$; and otherwise continued until the maximum sample size (N) was reached.

We assigned a point mass density with $p = 0.2$ under the null hypothesis H_0 and test

various prior densities under the alternative H_1 . The prior densities under H_1 include the standard Beta prior, the truncated Beta prior, the Uniform prior, and the iMOM prior. The first three belong to the class of local priors, while iMOM is a nonlocal prior. The iMOM and Beta priors ascribed to H_1 had the mode of 0.4. The strength of the prior was reflected by prior ESS. We considered both a weakly informative prior with ESS=5 and a strongly informative prior with ESS=20. For the weakly informative prior, we specified the nonlocal prior iMOM as $\pi(p; k = 0.325, \nu = 0.65, \tau = 0.7029)_{[0.2, 1]}$, and the local prior Beta(2.2, 2.8), based on the algorithm in section 6.3.3. To mimic the behavior of iMOM, i.e., assigning non-zero density at the parameter space with $p < 0.2$, we included a truncated beta prior Beta(2.2, 2.8) $_{[0.2, 1]}$, denoted as tBeta, for comparison. We also note that this truncated beta prior is still a local prior, as it has positive density at $p = 0.2$. We then considered the most commonly used Uniform distribution to see how the informative prior would perform in comparison to the non-informative prior. The Uniform distribution was also truncated in (0.2, 1) and denoted as tUniform. For the strongly informative prior ESS=20, the priors were iMOM $\pi(p; k = 1.685, \nu = 3.37, \tau = 0.0467)$, Beta(8.2, 11.8), and Beta(8.2, 11.8) $_{[0.2, 1]}$.

We considered 13 scenarios with the true response rate ranging from 0 to 0.6, with an interval space of 0.05. For each scenario, we considered three different metrics for the comparison between the nonlocal iMOM and the three local Beta priors. (1) The probability of claiming efficacy, which was defined as the percentage of simulated trials claiming that the treatment is efficacious if the criteria $BF_{10} \geq 9$ is satisfied during or at the end of the trial. (2) The probability of claiming futility, which was defined as the percentage of simulated trials claiming that the treatment is futile if $BF_{01} \geq 9$ is satisfied. (3) The average number of patients treated in the 10,000 simulated trials for each scenario. In addition to the overall probability of claiming efficacy or futility, we also summarize such a probability at each interim look, as well as at the final analysis. If larger probabilities of making a correct decision are observed at earlier interim analyses, it indicates a more efficient design. Here, the correct decision refers to claiming efficacy under H_1 and claiming futility under H_0 .

6.4.2 Simulation results

(a) Monitoring with a weakly informative prior Figure 6.1 shows the operating characteristics of the TB monitoring procedure with different prior distributions assumed under H_1 . With the weakly informative prior, the use of an iMOM prior yields a comparable probability of claiming efficacy in comparison to the use of tUniform and Beta, while the use of tBeta has the

largest probability of claiming efficacy (panel a1). The iMOM prior outperforms local priors in terms of futility monitoring. As shown in panel a2, the design with the iMOM prior has a greater probability to claim futility when the treatment is indeed not better than the standard of care. For instance, when the true response rate is 0.2, there is a probability of 0.86 to claiming futility under the use of an iMOM prior. On the other hand, the use of tUniform and tBeta yields smaller probabilities of claiming futility. Notably, the use of a Beta prior causes abysmal behavior, i.e., the trial never terminates for futility, regardless of the true response rate. Such a finding showcases the shortcoming of using a local prior in test-based monitoring.

In terms of average sample size, the design using the iMOM prior requires fewer patients on average. As shown in panel a3, when the true response rate is 0.2, the average number of patients enrolled is only 23 under the use of an iMOM prior, while this number is 32, 39, and 50 when tUniform, tBeta, and Beta are used, respectively.

(b) Monitoring with a strongly informative prior When a strongly informative prior (e.g., ESS=20) is used, we find that the use of iMOM, tBeta, and Beta priors have similar probabilities of claiming efficacy, and all outperform the use of the tUniform prior in terms of the probability of claiming efficacy when the true response rate is greater than 0.2. This indicates that when the prior is congruent with the data, the larger the ESS, the greater the power to make a correct decision. Like that with a weakly informative iMOM prior, the design with a strongly informative iMOM prior also has the greatest probability to claim futility when the treatment is indeed not better than the standard of care. The use of Beta with ESS=20 greatly increases the probability of claiming futility when the treatment is not promising (panel b2), compared to panel a2. Despite such an improvement, the design using a Beta prior still has a much lower probability of claiming futility than found with iMOM. As shown, with a Beta prior, when the true response rate is 0.2, the probability to claim futility is 0.55, much smaller than found with an iMOM prior (0.89). In terms of the number of patients treated, the use of a strongly informative iMOM prior still performs the best among the considered priors.

(c) Probability of making a correct decision at each interim The advantage of the iMOM prior is more prominent, as demonstrated by the probability of making correct decisions at each interim analysis. For example, the design with the iMOM prior stops much earlier for claiming futility when the response rate is 0.2 under both weakly and strongly informative priors (Figures 6.2 and A6.4, respectively). When H_1 is true (i.e., the response rate is 0.4), the behavior of the iMOM

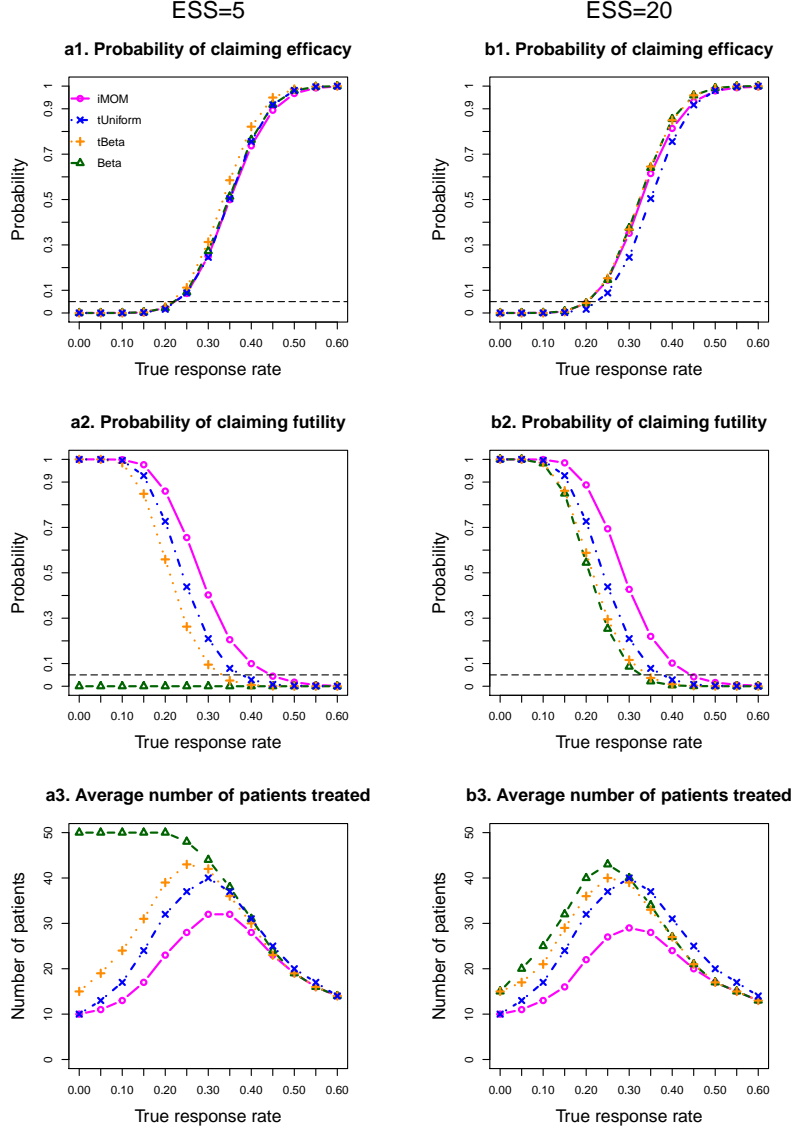


Figure 6.1: Efficacy and futility sequential monitoring using Bayes factor with weakly informative prior (ESS=5, left panels) and strongly informative prior (ESS=20, right panels). When ESS=5, the iMOM has parameters $k = 0.325$, $\nu = 0.65$, $\tau = 0.7029$ and the local prior is Beta(2.2, 2.8). When ESS=20, the iMOM has parameters $k = 1.685$, $\nu = 3.37$, $\tau = 0.0467$ and the local prior is Beta(8.2, 11.8). The maximum sample size is 50.

prior is comparable to that of tBeta and Beta in terms of stopping the trial earlier for efficacy.

In Supplementary Material, we detail the sensitivity analysis we conducted to assess the operating characteristics of the design using different priors with maximum sample size (N) varying from 20 to 200 and a prior effective sample size that varies from 5 to 20 (section A6.0.4). In summary, the use of iMOM is more efficient than the use of the other local priors, due to the fact that iMOM yields a smaller average sample size with a comparatively probability of claiming efficacy, when

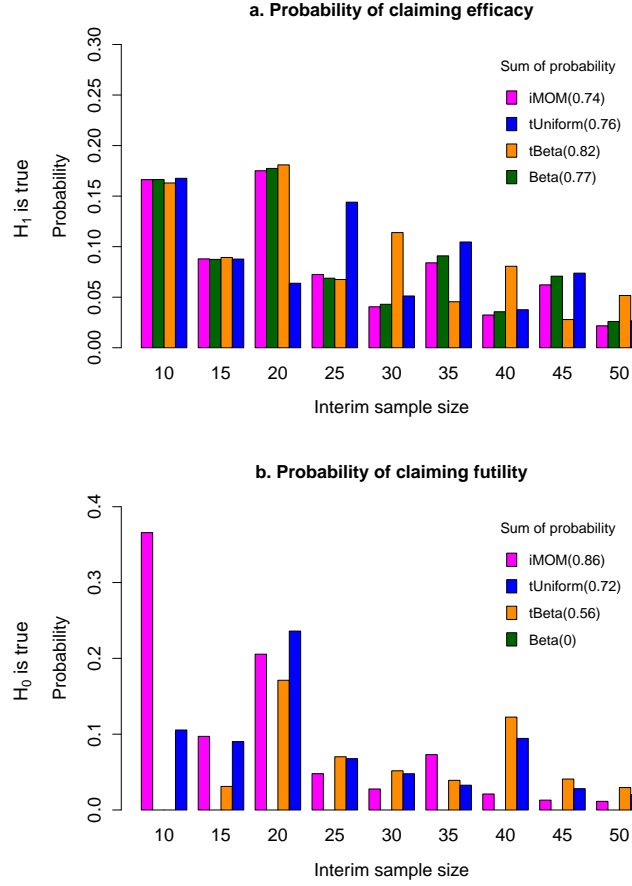


Figure 6.2: Probabilities of claiming efficacy when H_1 is true (i.e., panel a), and probabilities of claiming futility when H_0 is true (i.e., panel b) at each interim time and final analysis when maximum sample size $N = 50$ and prior effective sample size is 5 . The numbers in parentheses represent the overall probabilities of claiming futility or efficacy in the trial.

the treatment is truly effective, and a higher probability of claiming futility, when the treatment is truly ineffective.

6.5 Software application

To facilitate the use of an iMOM prior for TB monitoring in phase II trials, we developed a user-friendly Shiny application. Figure 6.3 shows the app interface of the **Trial setting** for both efficacy and futility TB monitoring using an iMOM prior. Users can use the app to get full stopping boundaries, run simulations, and download a formatted trial protocol template.

To obtain full stopping boundaries, users first set up the **Trial setting** by (1) entering the response rate under null and alternative hypothesis, respectively; (2) determining stopping rules

Trial Setting
Simulation
Protocol
User guide

Hypothesis

Null hypothesis (H_0): response rate is

Alternative hypothesis (H_1): response rate is

Interim monitoring

Purpose of monitoring

☐ Efficacy monitoring only
☐ Futility monitoring only
☒ Both efficacy and futility monitoring

Stop the trial for futility if BF_{01} (Bayes factor of H_0 over H_1) \geq

Stop the trial for Efficacy if BF_{10} (Bayes factor of H_1 over H_0) \geq

Sample size

Cohort size for monitoring the trial

Minimum number of patients to enroll for first monitoring

Maximum number of patients

Prior information

Prior effective sample size

Calculating stopping boundaries
Edit

Figure 6.3: Shiny app interface of the TB monitoring procedure based on the iMOM prior.

under *Interim monitoring*; (3) supplying cohort size, maximum number of patients, and minimum number of patients to enroll before the first monitoring under *Sample size*; and (4) providing the prior effective sample under *Prior information*. After the setup, click on the “calculate stopping boundaries” button . The full stopping boundaries will be shown and can be downloaded as csv, excel, or pdf files, as shown in Figure A6.8. For the example with settings presented in Figure 6.3, the full stopping boundaries are as shown in Table 6.1.

Number of patients	Stop for futility if number of responses is	Continue the trial if number of responses is	Stop for efficacy if number of responses is
10	0-1	2-5	6-10
15	0-2	3-6	7-15
20	0-4	5-8	9-20
25	0-5	6-10	11-25
30	0-6	7-11	12-30
35	0-8	9-13	14-35
40	0-9	10-14	15-40
45	0-11	12-16	17-45
50	0-12	13-17	18-50

Table 6.1: Full stopping boundaries for efficacy and futility monitoring using Bayes factor. The null response rate is 0.2 and the alternative response rate is 0.4. The trial stops for futility if $BF_{01} > 9$ and for efficacy if $BF_{10} > 9$. The prior effective sample size for the iMOM prior is 10. Patients enter the trial with a cohort size of five and the first interim is conducted after ten patients are treated.

Suppose at current interim analysis, where 20 patients are treated and seven responses are observed, the trial should continue to enroll more patients based on the stopping boundaries provided.

To run the simulation for various scenarios, users can use the **Simulation** tab after calculating stopping boundaries under **Trial setting**. As shown in Figure A6.9, enter the true response rate and the desired number of simulated trials for each scenario, click on the “Simulate” button. The simulation results will be shown on the right side of the app under the tab *Operating characteristics*. If users want to save their current parameter input for future use, they can easily do that under the *Summary of parameter input* next to *Operating characteristics*. The simulation results for our example are shown in Table 6.2. The results summarize the probabilities of claiming efficacy and futility under each scenario, the average number of patients treated, and some quantiles that may be of interest.

To generate the trial protocol, click on **Protocol** and a protocol template with full stopping boundaries and simulation results can be downloaded as either an html or word file. The protocol template for the current example can be seen in Figure A6.10. We also provide an extensive user guide to illustrate the methods and how to use the app. This information is available under the **User guide** tab.

True response	Pr(claim efficacy)	Pr(claim futility)	Patients allocation*
0.20	0.0282	0.8910	22(10,10,20,30,45)
0.30	0.3210	0.4154	31(10,15,30,50,50)
0.40	0.8014	0.0976	25(10,15,20,35,50)
0.50	0.9796	0.0126	17(10,10,15,20,30)
0.60	0.9984	0.0014	13(10,10,10,15,15)

*Patients allocation: average number of patients and quantiles (10%, 25%, 50%, 75%, 90%)

Table 6.2: Operating characteristics for a phase II study with sequential monitoring for both efficacy and futility using Bayes factor with iMOM prior. The null response rate is 0.2 and the alternative response rate is 0.4. The trial stops for futility if $BF_{01} > 9$ and for efficacy if $BF_{10} > 9$. The prior effective sample size for the iMOM prior is 10. Patients enter the trial with a cohort size of five. The first monitoring is conducted after ten patients are treated and subsequent interim monitoring is carried out after every five patients. The maximum number of patients enrolled is 50.

6.6 Summary

In this study, we first investigated the connection between the TB and PB approaches. In particular, we showed that TB sequential monitoring is essentially a special case of PB monitoring with a sample-size-dependent stopping cutoff. In the second part of this chapter, we extensively examined the effect of local and nonlocal priors on TB monitoring through simulation studies under various settings with different maximum sample sizes and prior effective sample sizes. When the treatment is indeed efficacious, using iMOM yields comparable probability of claiming efficacy in comparison to the use of local priors, and the average number of patients are also comparable across the use of the different priors. When the treatment is futile, the use of an iMOM prior has a much greater probability to claim futility. The TB monitoring with iMOM also has a larger probability of claiming futility at earlier interim analyses, which in turn results in a much smaller number of patients enrolled, reducing the number of patients exposed to futile treatment.

In addition, we also noted that a phase II trial with an iMOM prior for TB monitoring requires fewer patients to claim futility than to claim efficacy. This finding is consistent with what was observed in the original reference [46] in which it was shown that the Bayes factor converges faster to the true null hypothesis than to the true alternative hypothesis. Another important observation is that the use of a Beta prior never results in claiming of futility when ESS is small, even when the treatment is actually futile. With the increase in ESS, the probability of claiming futility using a Beta prior will greatly increase, although it is still smaller than that of using an iMOM prior.

In summation, we recommend the use of the iMOM prior for TB monitoring with Bayes

factor. To facilitate the use of an iMOM prior in trial monitoring, we developed a user-friendly web app, which is freely available at www.trialdesign.org.

A Appendix

A6.0.1 Interpretation of Bayes factor as the strength of evidence

Under the Bayesian hypothesis testing framework, a decision to reject the null hypothesis occurs only when the Bayes factor in favor of the alternative hypothesis exceeds a prespecified evidence level. Thus, it can quantify evidence in favor of the null hypothesis when the null hypothesis is not rejected [55]. Table A6.1 provides the interpretation of the strength of evidence in terms of both BF_{10} and $\log(BF_{10})$.

Table A6.1: Interpretation of Bayes factor(BF_{10}) and its log scale as the strength of evidence for H_1 over H_0 [47]

BF_{10}	$\log(BF_{10})$	Strength of evidence for H_1
<1:1	<0	Support H_0
1:1 to 3:1	0 to 1	Barely worth mentioning
3:1 to 20:1	1 to 3	Substantial
20:1 to 150:1	3 to 5	Strong
>150:1	>5	Very strong

A6.0.2 iMOM density with varying parameters

We show the iMOM density as a function of (k, τ, ν) in Figure A6.1. It is noted that an iMOM with larger values of k and τ has a wider null region and a heavier tail. In contrast, an iMOM with a larger value of ν tends to have lighter tails.

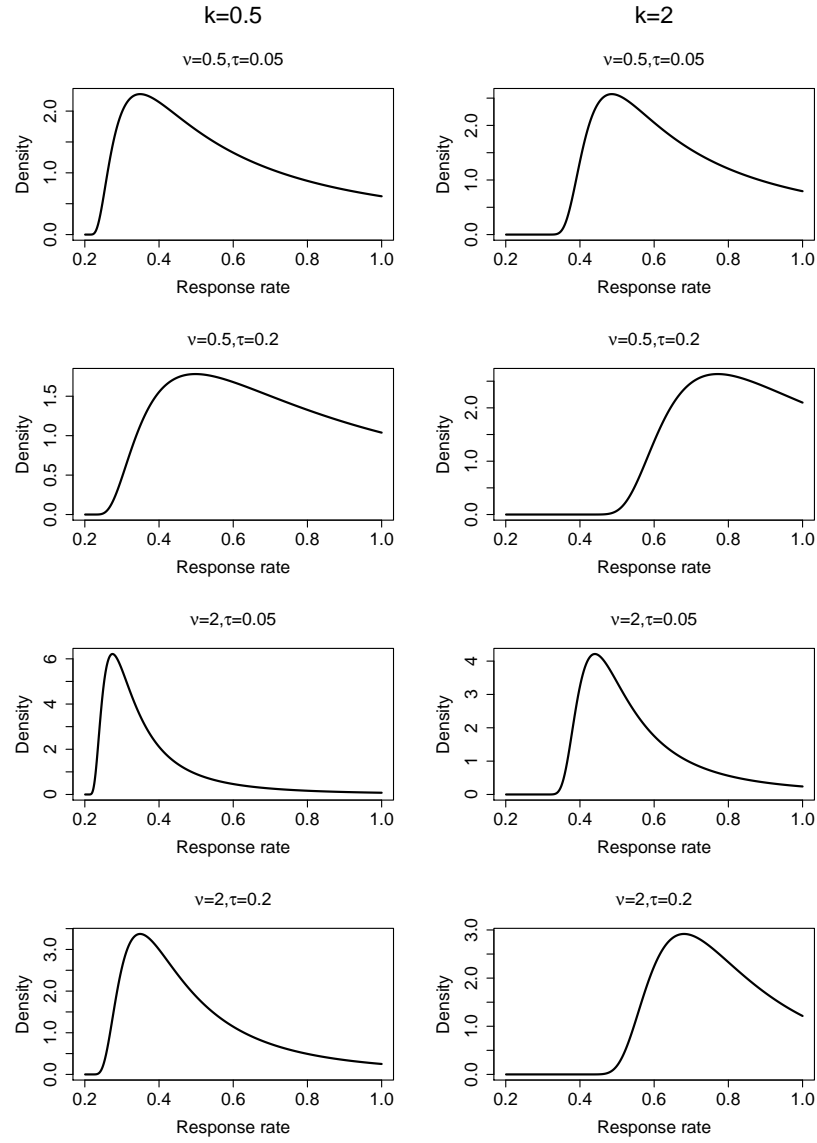


Figure A6.1: iMOM density with a null value at 0.2 under different parameter (k, τ, ν) settings.

A6.0.3 The relative robustness of iMOM versus Beta

We show here, under the setting described in [25] that iMOM has similar performance to the robust Cauchy prior (Figure A6.2).

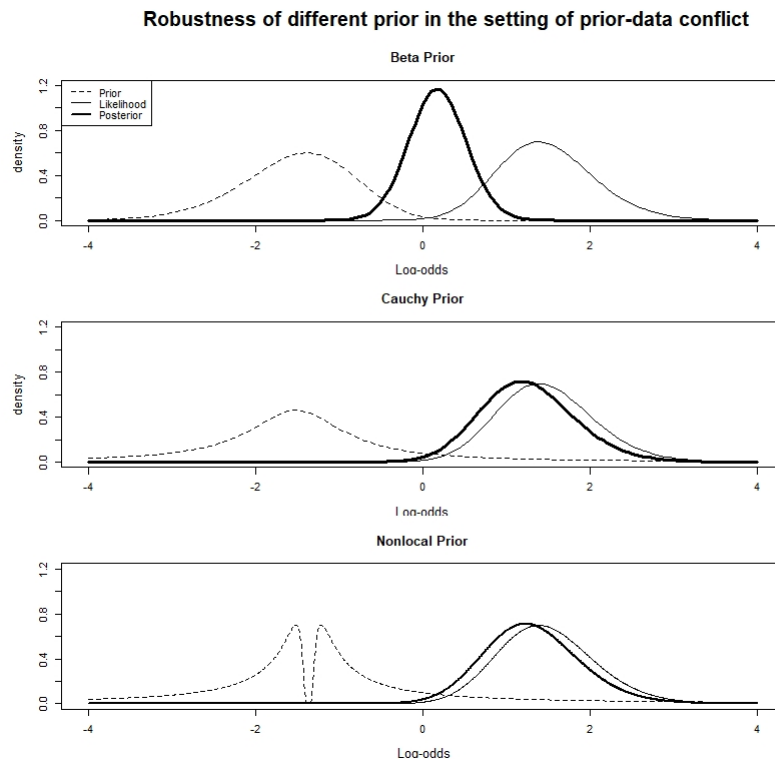


Figure A6.2: Plot of prior, likelihood, and posterior distribution for the setting described in [25]. All the prior distributions have the same mode. The prior mean response rate is 0.2, corresponds to a prior log(odds) of -1.39 , but the observed response rate is 0.8 with a sample size of 20.

A6.0.4 Sensitivity analysis

We conducted a sensitivity analysis to assess the operating characteristics of the design using different priors with maximum sample size (N) varying from 20 to 200 and prior effective sample size varying from 5 to 20. We show the results at $N = 100$ and ESS= (5, 20) in Figure A6.3. As expected, the rise in both N and ESS escalates the probability of claiming efficacy when the treatment is promising, and increases the probability of claiming futility when the treatment is futile. Under all settings, the use of an iMOM prior requires much fewer patients (panels a3 and b3 in Figure A6.3). This is because the trial always stops at earlier interim analyses for futility (e.g., panel b in Figures A6.4-A6.7), thus requiring much fewer patients. We note that while under the Beta prior, ESS changes from 5 to 20 will increase the probability of claiming futility from 0 to

0.55 when $N = 50$ and from 0 to 0.79 when $N = 100$, but the stopping occurs at the later interim analyses compared to that under iMOM (panel b in Figures A6.4 and A6.7).

In terms of trial efficiency, we found that iMOM was more desirable for futility monitoring through all specifications of maximum sample size and prior effective sample size (e.g., Figures A6.4- A6.7). Given a fixed ESS (e.g., 5), the use of iMOM has a comparable probability of claiming efficacy in comparison to tBeta and beta when sample size is smaller, while it yields slightly smaller probability of claiming efficacy when sample size is larger. For instance, when sample size decreases to 25 (Figure A6.5), iMOM, tBeta, and Beta have the same probability of claiming efficacy (0.5), which is slighter larger than that under tUniform (0.46); when sample size increases to 100 (Figure A6.6), we note that iMOM has a smaller probability of claiming efficacy: 0.06%, 0.08%, 0.09% smaller than tUniform, tBeta, and Beta, respectively. Given a fixed sample size (e.g., 50), the larger the ESS, the smaller the difference between using iMOM and using tBeta in terms of the probability of claiming efficacy. For example, when ESS=5 (Figure 6.2), the difference between iMOM and tBeta is 0.08, but when ESS=20 (Figure A6.4), the difference is slightly smaller (0.03).

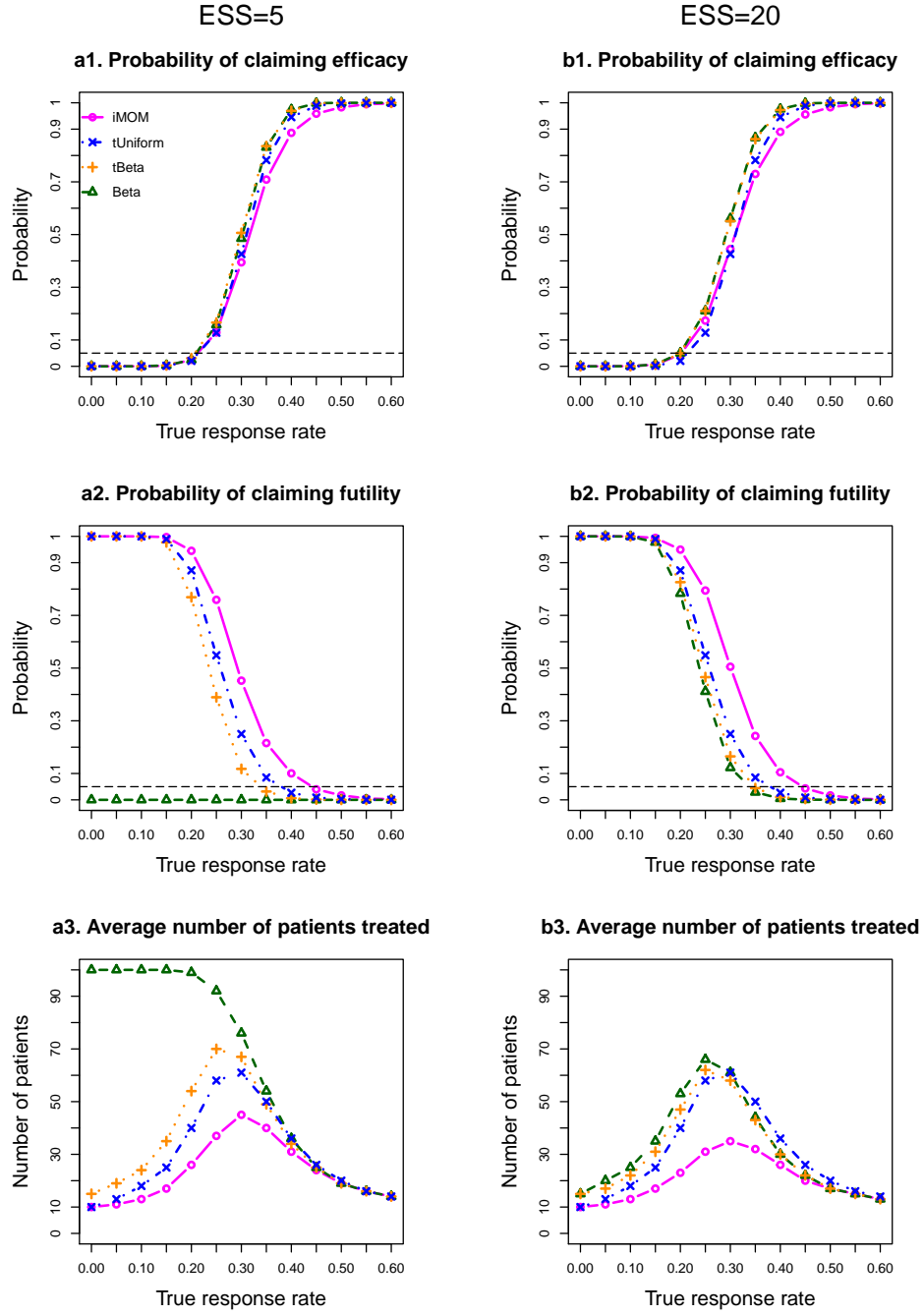


Figure A6.3: Efficacy and futility sequential monitoring using Bayes factor with weakly informative prior distributions (ESS=5) and a strongly informative prior (ESS=20). When ESS=5, the iMOM has parameters $k = 0.325$, $\nu = 0.65$, $\tau = 0.7029$ and the local prior is Beta(2.2, 2.8). When ESS=20, the iMOM has parameters $k = 1.685$, $\nu = 3.37$, $\tau = 0.0467$ and the local prior is Beta(8.2, 11.8). The maximum sample size is 100.

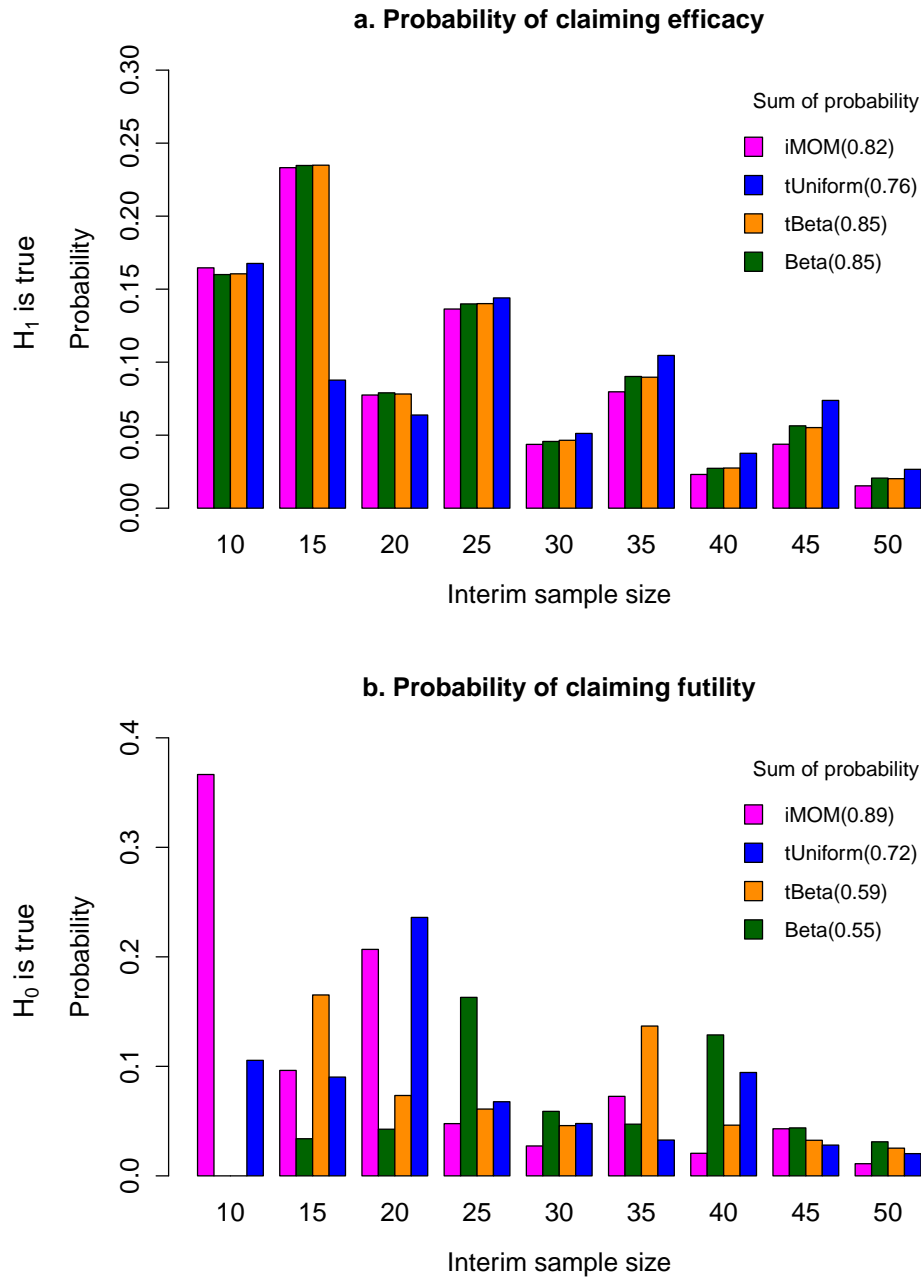


Figure A6.4: Probability of claiming efficacy under H_1 and probability of claiming futility under H_0 at each interim or final analysis when maximum sample size $N = 50$ and prior effective sample size is 20 . The numbers in parentheses represent the overall probabilities of claiming futility or efficacy in the trial.

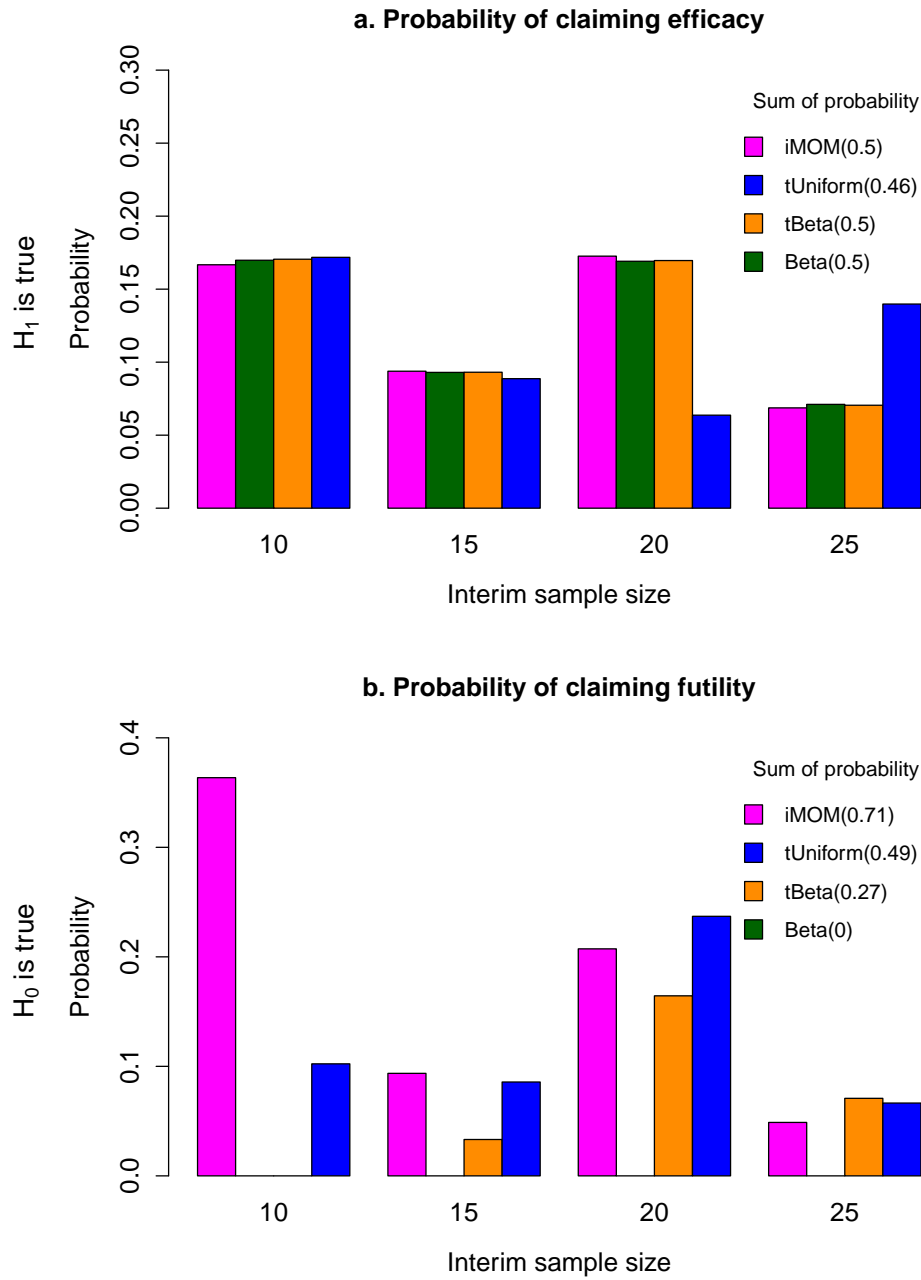


Figure A6.5: Probability of claiming efficacy under H_1 and probability of claiming futility under H_0 at each interim or final analysis when maximum sample size $\mathbf{N} = 25$ and prior effective sample size is $\mathbf{5}$. The numbers in parentheses represent the overall probabilities of claiming futility or efficacy in the trial.

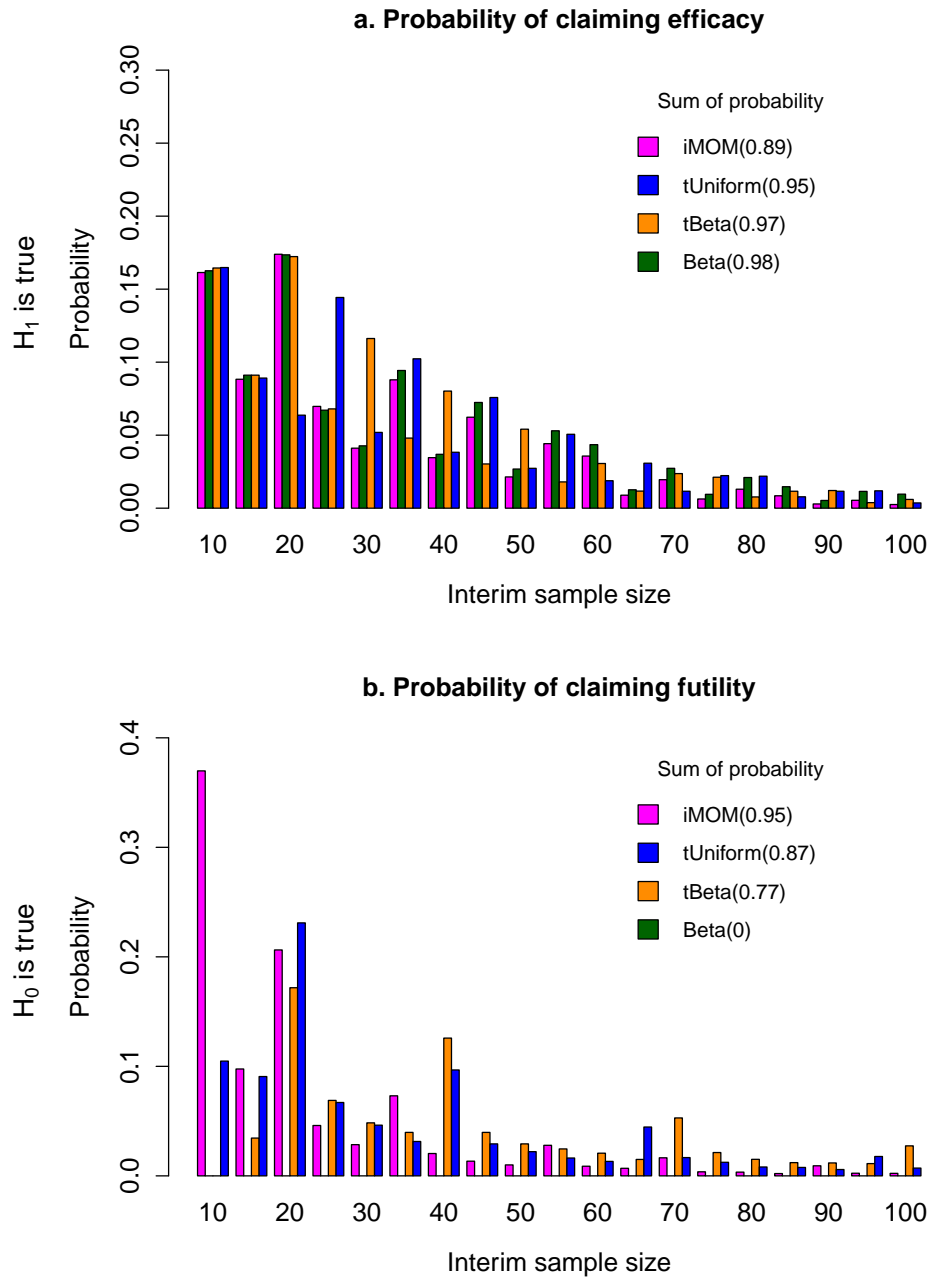


Figure A6.6: Probability of claiming efficacy under H_1 and probability of claiming futility under H_0 at each interim or final analysis when maximum sample size $N = 100$ and prior effective sample size is 5 . The numbers in parentheses represent the overall probabilities of claiming futility or efficacy in the trial.

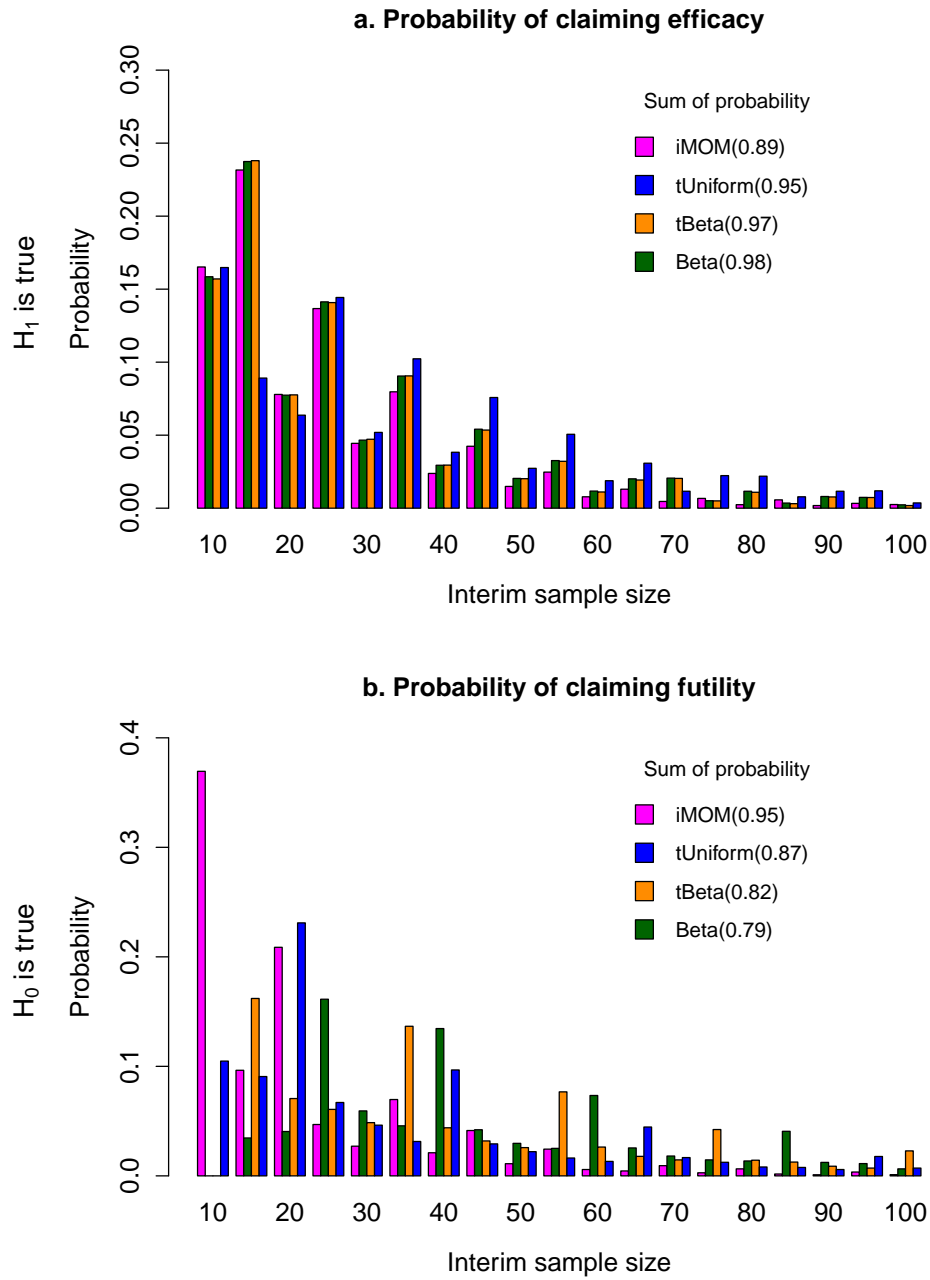


Figure A6.7: Probability of claiming efficacy under H_1 and probability of claiming futility under H_0 at each interim or final analysis when maximum sample size $N = 100$ and prior effective sample size is 20 . The numbers in parentheses represent the overall probabilities of claiming futility or efficacy in the trial.

A6.0.5 Output for the shiny application example

In this section, we show examples of the typical outputs for TB monitoring using iMOM prior with the web-based app available at www.trialdesign.org. Figure A6.8 shows the full stopping boundaries and the corresponding iMOM prior given trial parameters set in Figure 6.3. Figure A6.9 displays the output for simulations when a trial is planned. And Figure A6.10 showcases the formatted protocol for the example provided in the main text.

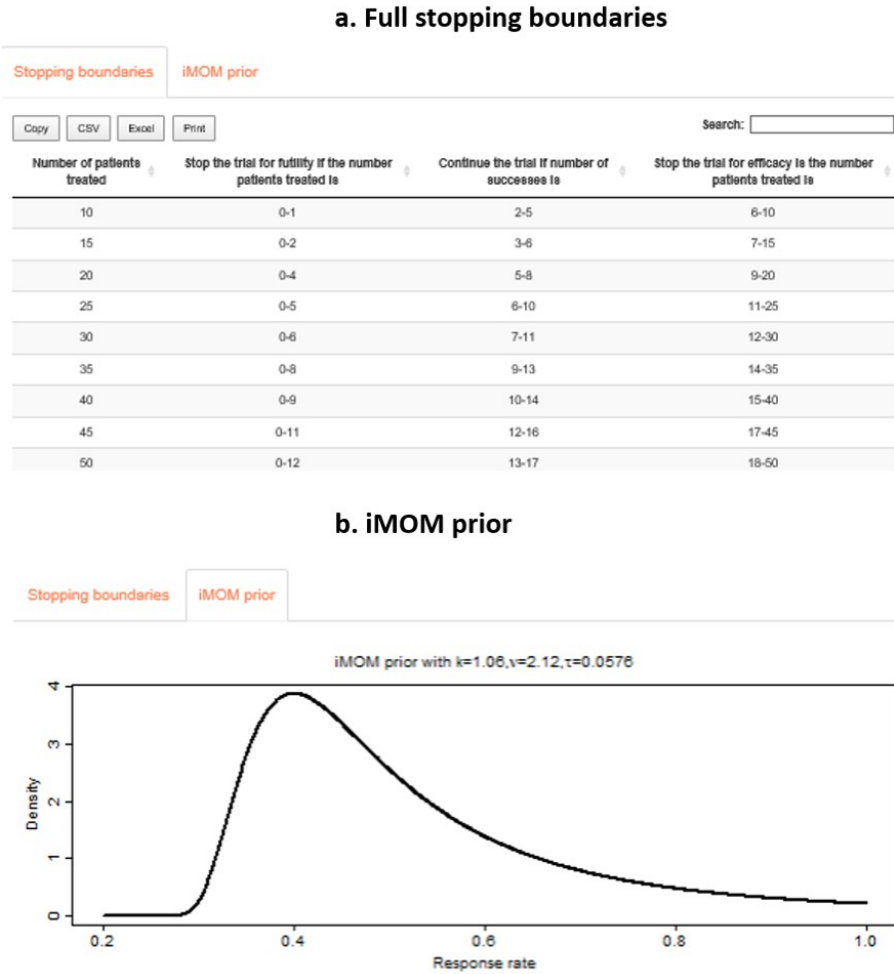


Figure A6.8: Full stopping boundaries and iMOM prior determined by the input under **Trial setting** in the Shiny app, as shown in Figure 6.3.

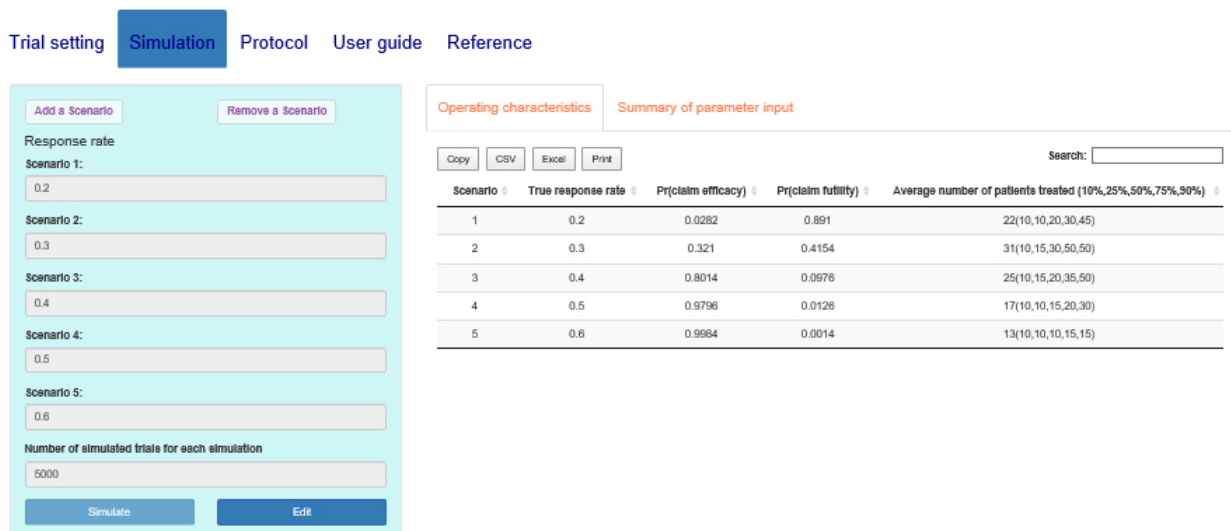


Figure A6.9: Example of simulation output.

Template for Protocol Preparation

Statistical Design

We use a single arm phase II trial to assess treatment effect of *name of the treatment* in patients with *disease X*. The primary endpoint is objective response rate, defined as complete response rate plus partial response rate. We will compare the response rate of the treatment in the trial to that of the standard of care.

Based on historical data, the response rate for the standard of care is 0.2, which is our null hypothesis (H_0). We expect that the treatment would have a response rate of 0.4, which is assumed under the alternative hypothesis (H_1). The Bayesian hypothesis test-based design (Johnson & Cook, 2009; Zhou et al., 2019) will be used for decision making. We assume that the sample distribution of number of responses follows a binomial distribution, and use an inverse moment density as the prior (Johnson 2010) for the response rate under the alternative hypothesis. The prior effective sample size is 10. As data accumulates, we calculate the Bayes factor of H_1 over H_0 (denoted as BF_{10}), or Bayes factor of H_0 over H_1 (denoted as BF_{01}).

Stopping rules and boundaries

The cohortsize for monitoring the trial is 5. A minimum of 10 patients will be enrolled for the first interim monitoring. A maximum of 50 patients will be accrued.

We implement two stopping rules during the trial:

- (1) Stop the trial for futility if the Bayes factor $BF_{01} \geq 9$.
- (2) Stop the trial for efficacy if the Bayes factor $BF_{10} \geq 9$.

The trial will stop if the number of patients treated reaches the maximum sample size.

The full stopping boundary is shown in Table 1.

Table 1. Full stopping boundaries

Number of patients treated	Stop the trial for futility if the number patients treated is	Continue the trial if number of successes is	Stop the trial for efficacy if the number patients treated is
10	0-1	2-5	6-10
15	0-2	3-6	7-15
20	0-4	5-8	9-20
25	0-5	6-10	11-25
30	0-6	7-11	12-30
35	0-8	9-13	14-35

1

40	0-9	10-14	15-40
45	0-11	12-16	17-45
50	0-12	13-17	18-50

Operating Characteristics

The operating characteristics of the design were based on 5000 simulated trials for each scenarios. The results were produced using the Shiny app BFMonitoring version V1.0.0.0.

Table 2. Operating characteristics

Scenario	True response rate	Pr(claim efficacy)	Pr(claim futility)	Average number of patients treated (10%,25%,50%,75%,90%)
1	0.2	0.0282	0.891	22(10,10,20,30,45)
2	0.3	0.321	0.4154	31(10,15,30,50,50)
3	0.4	0.8014	0.0976	25(10,15,20,35,50)
4	0.5	0.9796	0.0126	17(10,10,15,20,30)
5	0.6	0.9984	0.0014	13(10,10,10,15,15)

Reference

Johnson, V. E., & Cook, J. D. (2009). Bayesian design of single-arm phase II clinical trials with continuous monitoring. *Clinical Trials*, 6(3), 217-226.

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2), 143-170.

Zhou, Y., Lin, R., & Lee, J. J. (2019). The Use of Local and Nonlocal Priors in Bayesian Test-Based Monitoring Designs for Phase II Clinical Trials (*Submitted*)

Zhou, Y., Lin, R., & Lee, J. J. (2019). Single arm phase II monitoring using Bayes factor and iMOM prior for binary outcome: Version 1.0.0.0. Web, available at <http://www.trialdesign.org>.

2

Figure A6.10: An example for the template of protocol.

CHAPTER 7

Conclusion

In Chapter 2, we reviewed the the designs available for Phase I clinical trials, and elucidated the similarity and difference between the recently proposed algorithm-based i3+3 design and the model-assisted BOIN design. We showed that the i3+3 design adopted very similar decision rules as the BOIN design, but because of taking the algorithm-based design framework, it suffers from a series of scientific and logical deficiencies inherited in algorithm-based designs. We further showed that the “de-escalation modification rule” used by i3+3 to differentiate it from BOIN does not improve, but often impair the design performance. Simulation study shows that compared to the i3+3 design, the BOIN design has better accuracy to correctly identify the MTD and is safer with lower risk of overdosing patients. In addition, BOIN is also simpler, more transparent and has been widely validated in a variety of clinical trials, thus provides a better choice for Phase I oncology trials.

Although BOIN has been well-validated and used in a wide range of Phase I clinical trials, it has been perceived by many that it, as a model-assisted design, cannot incorporate historical information or real-world evidence into trial planning and conduct. Similar concerns are true for other model-assisted designs, such as mTPI and Keyboard/mTPI-2. To improve the efficiency of the model-assisted Phase I designs, in Chapter 3, we proposed a systematic way to incorporate prior estimated DLT probabilities obtained from historical data, external trials, or real-world evidence into the designs. The proposed method takes a similar approach as the CRM skeleton and combines it with the concept of prior effective sample size to incorporate informative prior information into the model-assisted designs, including BOIN and keyboard. The resulting designs are referred to as iBOIN and iKeyboard, respectively. The approach is readily applicable to mTPI. The reason why we did not include it here is that this design has a severe problem of overdosing patients and thus not a good alternative for Phase I clinical trials. A robust prior is also proposed to accommodate

the case where clinicians are not very certain about the prior estimate of DLT probabilities for the doses under investigation. Additionally, in this chapter, we also quantified the effective sample size induced by the normal distribution on the hyper-parameter in CRM, using a more straightforward approach. CRM using informative prior is denoted as iCRM. Our simulation results demonstrate that when prior MTD is correctly specified, all the informative designs greatly improve both the percentage of correctly selecting the true MTD and the percentage of assigning patients to the MTD, with the largest improvement observed in iBOIN. Both iCRM and iKeyboard are riskier than iBOIN. The iKeyboard is also the most sensitive design to prior specification. We recommended iBOIN for Phase I clinical trials when good prior information is available, due to the simplicity and superior performance of the design. iBOIN with robust prior or standard BOIN should be used if it is anticipated that the prior may not well approximate the true DLT probability curve.

In Chapter 4, we proposed the U-BOIN, a seamless Phase I/II model-assisted design, to identify the optimal biological dose (OBD) for targeted and immunotherapy trials. The U-BOIN design accounts for the efficacy-toxicity trade-off using a utility function. Unlike most existing Phase I/II designs, which require complicated real-time model fitting and estimation to make dose assignment decisions, the U-BOIN is simple and easy to implement. The dose assignment rules of the U-BOIN can be pre-tabulated in decision tables and included in the trial protocol before the onset of a trial. To conduct the trial, no complicated calculation is needed. The investigator can simply use the decision tables to make the decision of dose escalation/de-escalation. Simulation studies show that compared to a more complicated model-based Phase I/II design, the U-BOIN has higher accuracy to identify OBD and is more robust. Besides immunotherapy and targeted trials, U-BOIN also can be used for conventional cytotoxic agent trials. In such cases, both toxicity and efficacy typically increase with the dose, but may do so at different rates. It is likely that increasing the dose causes much higher toxicity, with limited efficacy gain. The idea of risk-benefit trade-off is still generally applicable here for use in most medical decisions in practice. Like most model-assisted designs, U-BOIN models efficacy and toxicity at each dose independently, whereas model-based Phase I/II designs (e.g., EffTox design) model efficacy and toxicity across all doses, through imposing a parametric dose-efficacy and -toxicity curve model. As a result, one may worry about the potential efficiency loss for U-BOIN. Our numerical study and previous studies show that, for the purpose of dose finding, the efficiency loss caused by using only local data (in model-assisted designs, such as U-BOIN) is minimal or negligible. This is because, to make correct decisions of dose assignment and selection, we only need to correctly estimate the rank of utility across the

doses. A slightly more variability on the estimate of the utility has no or negligible impact on the performance of the design.

In addition to our extensions of the BOIN to incorporate informative prior (iBOIN) and to identify OBD by measuring dose desirability using utility function (U-BOIN), there are many other designs that have built upon the BOIN design to address different challenges in Phase I clinical trials. For instance, it has been extended to address the issues brought by fast accrual, delayed outcome, and evaluation of drug combinations. Keyboard has been shown to have comparable operating characteristics to BOIN in numerous studies and it also has been extended to address delayed toxicity problems and to identify MTD in drug-combination trials. However, the use of these designs is limited due to the lack of reliable, robust, and user-friendly software. To facilitate the use of the model-assisted designs for designing and conducting Phase I clinical trials and make it more accessible for clinicians who may not have expertise in statistical programming, we developed a software platform for BOIN designs (BOIN suite) and keyboard designs (Keyboard suite), respectively, in Chapter 5. The BOIN suites include designs for single-agent trials using non-informative prior (standard BOIN) or using informative prior derived from historical data (iBOIN); for single-agent trials with fast accrual or delayed toxicity (TITE-BOIN); for drug-combination (BOIN COMB) to find a single MTD or a MTD contour; and for utility-based BOIN design (U-BOIN) to find OBD for immunotheapy and targeted therapy trials. Keyboard suite includes designs for single-agent trials without delayed toxicity (KEYBOARD [105]) or with delayed toxicity (TITE-KEYBOARD [57]); and for drug-combination trials to find a single MTD (KEYBOARD COMB [80]). The strengths of the BOIN suite and Keyboard suite are that they (1) run under different operating systems; (2) do not require any installation; (3) do not require statistical programming skills to use; (4) provide comprehensive user help files; and (5) are regularly maintained and free to use. Another unique feature of the two software suite is that all the designs are model-assisted and thus provide simple-to-use decision rules like that in 3+3, which is easy for practitioners familiar with 3+3 to adopt these new adaptive designs.

In Chapter 6, we first studied the connection between the Bayesian hypothesis test based (TB) and posterior/predictive probability based (PB) approaches for sequential monitoring in Phase II clinical trials. In particular, we showed that the TB sequential monitoring (using Bayes factor) is essentially a special case of the PB monitoring with a sample-size-dependent stopping cutoff. In the second part of this chapter, we extensively examined the effect of local and nonlocal prior on TB monitoring through simulation studies under various settings with different maximum sample

sizes and prior effective sample sizes. The TB monitoring with nonlocal prior (e.g., iMOM prior) has larger probability of claiming futility at earlier interim analyses when the treatment under investigation is indeed futile, which in turn result in much smaller number of patients enrolled, reducing the number of patients exposed to futile treatment. Additionally, we also noted that a Phase II trial with iMOM prior for TB monitoring requires fewer patients to claim futility than to claim efficacy. This finding is consistent with what observed in the original reference, in which it was shown that the Bayes factor converges faster to the true null hypothesis than to the true alternative hypothesis. Another important observation is that the use of local prior (e.g., Beta prior) never results in claiming of futility when ESS is small, even when the treatment is actually futile. With the increase in ESS, the probability of claiming futility using Beta prior will greatly increase, although it is still smaller than that using nonlocal prior. For the above reasons, we recommended the use of iMOM prior for TB monitoring with Bayes factor. To facilitate the use of iMOM prior in trial monitoring, we also developed a user-friendly web-based app, which is freely available at www.trialdesign.org.

To sum up, chapters 2-6 contain five projects that make significant contributions to drug development, as these projects advance the research of early phase clinical trial designs by providing novel model-assisted designs and facilitate a broader use of efficient model-assisted designs in early phase clinical trials by constructing user-friendly apps.

Bibliography

- [1] Al-Atrash, G. (2018). Nivolumab and ipilimumab after donor stem cell transplant in treating participants with high risk refractory or relapsed acute myeloid leukemia. *ClinicalTrials.gov Identifier: NCT03600155*.
- [2] Babb, J., Rogatko, A., and Zacks, S. (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine*, 17(10):1103–1120.
- [3] Berry, D. A. (2003). Statistical innovations in cancer research. *Cancer Medicine*, 6:465–478.
- [4] Berry, S. M., Carlin, B. P., Lee, J. J., and Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. CRC press.
- [5] Braun, T. M. (2002a). The bivariate continual reassessment method: extending the crm to phase I trials of two competing outcomes. *Controlled Clinical Trials*, 23(3):240 – 256.
- [6] Braun, T. M. (2002b). The bivariate continual reassessment method: extending the crm to phase I trials of two competing outcomes. *Controlled Clinical Trials*, 23(3):240–256.
- [7] Bril, G., Dykstra, R., Pillers, C., and Robertson, T. (1984). Algorithm as 206: isotonic regression in two independent variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(3):352–357.
- [8] Cai, C., Liu, S., and Yuan, Y. (2014). A bayesian design for phase II clinical trials with delayed responses based on multiple imputation. *Statistics in Medicine*, 33(23):4017–4028.
- [9] Chang, M. N., Therneau, T. M., Wieand, H. S., and Cha, S. S. (1987). Designs for group sequential phase II clinical trials. *Biometrics*, pages 865–874.
- [10] Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J., et al. (2017). Shiny: web application framework for R. *R Package Version*, 1(5).
- [11] Chen, T. T. (1997). Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*, 16(23):2701–2711.
- [12] Cheung, Y. K. (2011). *Dose finding by the continual reassessment method*. CRC Press.
- [13] Cheung, Y. K. and Chappell, R. (2000a). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*, 56(4):1177–1182.

- [14] Cheung, Y. K. and Chappell, R. (2000b). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*, 56(4):1177–1182.
- [15] Chevret, S. (2012). Bayesian adaptive clinical trials: a dream for statisticians only? *Statistics in Medicine*, 31(11-12):1002–1013.
- [16] Clertant, M. and O’Quigley, J. (2017). Semiparametric dose finding methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1487–1508.
- [17] Collins, J. M., Grieshaber, C. K., and Chabner, B. A. (1990). Pharmacologically guided phase I clinical trials based upon preclinical drug development. *JNCI: Journal of the National Cancer Institute*, 82(16):1321–1326.
- [18] Daud, A. I., Loo, K., Pauli, M. L., Sanchez-Rodriguez, R., Sandoval, P. M., Taravati, K., Tsai, K., Nosrati, A., Nardo, L., Alvarado, M. D., et al. (2016). Tumor immune profiling predicts response to anti-pd-1 therapy in human melanoma. *The Journal of Clinical Investigation*, 126(9):3447–3452.
- [19] Dixon, W. J. and Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association*, 43(241):109–126.
- [20] Dong, G. (2010). *A study of stagewise phase II and phase II/III designs for clinical trials*. PhD thesis, Rutgers University-Graduate School-New Brunswick.
- [21] Durham, S. D., Flournoy, N., and Rosenberger, W. F. (1997). A random walk rule for phase I clinical trials. *Biometrics*, 53:745–760.
- [22] FDA (2019). Interacting with the FDA on complex innovative trial designs for drugs and biological products draft guidance. *U.S. Food and Drug Administration*, pages 1–7.
- [23] Fleming, T. R. (1982). One-sample multiple testing procedure for phase II clinical trials. *Biometrics*, pages 143–151.
- [24] Food, U., Administration, D., et al. (2019). Submitting documents using real-world data and real-world evidence to fda for drugs and biologics: Guidance for industry: Draft guidance. *Rockville, MD: US Food and Drug Administration*.
- [25] Fúquene, J. A., Cook, J. D., Pericchi, L. R., et al. (2009). A case for robust bayesian priors with applications to clinical trials. *Bayesian Analysis*, 4(4):817–846.

- [26] Gehan, E. A. (1961). The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases*, 13(4):346–353.
- [27] Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., et al. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.
- [28] Gezmu, M. and Flournoy, N. (2006). Group up-and-down designs for dose-finding. *Journal of Statistical Planning and Inference*, 136(6):1749–1764.
- [29] Gilks, W. R., Best, N., and Tan, K. (1995). Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(4):455–472.
- [30] Guo, B. and Li, Y. (2014). Bayesian designs of phase II oncology trials to select maximum effective dose assuming monotonic dose-response relationship. *BMC medical research methodology*, 14(1):95.
- [31] Guo, W., Wang, S.-J., Yang, S., Lynn, H., and Ji, Y. (2017). A bayesian interval dose-finding design addressing cockham’s razor: mtpi-2. *Contemporary clinical trials*, 58:23–33.
- [32] Heitjan, D. F. (1997). Bayesian interim analysis of phase II cancer clinical trials. *Statistics in Medicine*, 16(16):1791–1802.
- [33] Hirakawa, A., Sato, H., Daimon, T., and Matsui, S. (2018). Dose finding in phase I cancer trials. In *Modern Dose-Finding Designs for Cancer phase I Trials: Drug Combinations and Molecularly Targeted Agents*, pages 1–7. Springer.
- [34] Hobbs, B. P., Chen, N., and Lee, J. J. (2018). Controlled multi-arm platform design using predictive probability. *Statistical Methods in Medical Research*, 27(1):65–78.
- [35] Houede, N., Thall, P. F., Nguyen, H., Paoletti, X., and Kramar, A. (2010). Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase I/II trials. *Bio-metrics*, 66(2):532–540.
- [36] Iasonos, A., Wilton, A. S., Riedel, E. R., Seshan, V. E., and Spriggs, D. R. (2008). A comprehensive comparison of the continual reassessment method to the standard 3+ 3 dose escalation scheme in phase I dose-finding studies. *Clinical Trials*, 5(5):465–477.

- [37] Ivanova, A., Flournoy, N., and Chung, Y. (2007). Cumulative cohort design for dose-finding. *Journal of Statistical Planning and Inference*, 137(7):2316–2327.
- [38] Ivanova, A., Montazer-Haghighi, A., Mohanty, S. G., and D Durham, S. (2003). Improved up-and-down designs for phase I trials. *Statistics in Medicine*, 22(1):69–82.
- [39] Jack Lee, J. and Chu, C. T. (2012). Bayesian clinical trials in action. *Statistics in Medicine*, 31(25):2955–2972.
- [40] Jacob, L., Uvarova, M., Boulet, S., Begaj, I., and Chevet, S. (2016). Evaluation of a multi-arm multi-stage bayesian design for phase II drug selection trials—an example in hemato-oncology. *BMC medical research methodology*, 16(1):67.
- [41] Jazaeri, A.A., Y. C. (2019). T cell immunotherapy for advanced ovarian cancer. *ClinicalTrials.gov Identifier: NCT03600155*.
- [42] Ji, Y., Liu, P., Li, Y., and Bekele, B. N. (2010). A modified toxicity probability interval method for dose-finding trials. *Clinical Trials*, 7(6):235–244.
- [43] Ji, Y. and Wang, S.-J. (2013). Modified toxicity probability interval design: a safer and more reliable method than the 3+ 3 design for practical phase I trials. *Journal of Clinical Oncology*, 31(14):1785.
- [44] Jin, I. H., Liu, S., Thall, P. F., and Yuan, Y. (2014). Using data augmentation to facilitate conduct of phase I-II clinical trials with delayed outcomes. *Journal of the American Statistical Association*, 109(506):525–536.
- [45] Johnson, V. E. and Cook, J. D. (2009). Bayesian design of single-arm phase II clinical trials with continuous monitoring. *Clinical Trials*, 6(3):217–226.
- [46] Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170.
- [47] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.
- [48] Le Tourneau, C., Lee, J. J., and Siu, L. L. (2009). Dose escalation methods in phase I cancer clinical trials. *Journal of the National Cancer Institute*, 101:708–720.

- [49] Lee, J. J. and Liu, D. D. (2008). A predictive probability design for phase II cancer clinical trials. *Clinical Trials*, 5(2):93–106.
- [50] Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.
- [51] Leonard, J.P., A. J. and Rutherford, S. (2018). Study of venetoclax plus DA-EPOCH-R for the treatment of aggressive B-Cell lymphomas (V+DA-EPOCH-R). *ClinicalTrials.gov Identifier: NCT03036904*.
- [52] Li, J. (2018). The safety, efficacy of anti-EGFR humanized monoclonal antibody combined with chemotherapy in advanced solid tumors (HLX07Ib/II). *ClinicalTrials.gov Identifier: NCT03577704*.
- [53] Li, Y. and Yuan, Y. (2020). PA-CRM: A continuous reassessment method for pediatric phase I oncology trials with concurrent adult trials. *Biometrics*, pages 1–10.
- [54] Lim, B. (2019). A phase II study of triple combination of atezolizumab + cobimetinib + eribulin (ACE) in patients with recurrent/metastatic inflammatory breast cancer. *ClinicalTrials.gov Identifier: NCT03202316*.
- [55] Lin, R. and Yin, G. (2015). Bayes factor and posterior probability: Complementary statistical evidence to p-value. *Contemporary Clinical Trials*, 44:33–35.
- [56] Lin, R. and Yin, G. (2017). Bayesian optimal interval design for dose finding in drug-combination trials. *Statistical Methods in Medical Research*, 26(5):2155–2167.
- [57] Lin, R. and Yuan, Y. (2019). Time-to-event model-assisted designs for dose-finding trials with delayed toxicity. *Biostatistics*.
- [58] Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*, volume 333. John Wiley & Sons.
- [59] Liu, J., Lin, Y., and Shih, W. J. (2010). On Simon’s two-stage design for single-arm phase IIa cancer clinical trials under beta-binomial distribution. *Statistics in Medicine*, 29(10):1084–1095.
- [60] Liu, M., Wang, S.-J., and Ji, Y. (2019). The i3+3 design for phase I clinical trials. *Journal of Biopharmaceutical Statistics*, pages 1–11.

- [61] Liu, S., Guo, B., and Yuan, Y. (2018). A bayesian phase I/II trial design for immunotherapy. *Journal of the American Statistical Association*, pages 1–12.
- [62] Liu, S. and Johnson, V. E. (2016). A robust bayesian dose-finding design for phase I/II clinical trials. *Biostatistics*, 17(2):249–263.
- [63] Liu, S., Pan, H., Xia, J., Huang, Q., and Yuan, Y. (2015). Bridging continual reassessment method for phase I clinical trials in different ethnic populations. *Statistics in Medicine*, 34(10):1681–1694.
- [64] Liu, S., Yin, G., and Yuan, Y. (2013). Bayesian data augmentation dose finding with continual reassessment method and delayed toxicity. *The Annals of Applied Statistics*, 7(4):1837.
- [65] Liu, S. and Yuan, Y. (2015). Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(3):507–523.
- [66] Loskog, A. (2019). Phase I/IIa trial evaluating safety of LOAd703, an armed oncolytic adenovirus for pancreatic cancer. *ClinicalTrials.gov Identifier: NCT02705196*.
- [67] Love, S. B., Brown, S., Weir, C. J., Harbron, C., Yap, C., Gaschler-Markefski, B., Matcham, J., Caffrey, L., McKevitt, C., Clive, S., et al. (2017). Embracing model-based designs for dose-finding trials. *British Journal of Cancer*, 117(3):332–339.
- [68] Morita, S. (2011). Application of the continual reassessment method to a phase I dose-finding trial in Japanese patients: East meets west. *Statistics in Medicine*, 30(17):2090–2097.
- [69] Morita, S., Thall, P. F., and Müller, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602.
- [70] Morita, S., Thall, P. F., and Müller, P. (2010). Evaluating the impact of prior assumptions in bayesian biostatistics. *Statistics in biosciences*, 2(1):1–17.
- [71] Murray, T. A., Thall, P. F., Yuan, Y., McAvoy, S., and Gomez, D. R. (2017). Robust treatment comparison based on utilities of semi-competing risks in non-small-cell lung cancer. *Journal of the American Statistical Association*, 112(517):11–23.
- [72] Murray, T. A., Yuan, Y., Thall, P. F., Elizondo, J. H., and Hofstetter, W. L. (2018). A utility-based design for randomized comparative trials with ordinal outcomes and prognostic subgroups. *Biometrics*.

- [73] Nakashima, K., Shimada, H., Ochiai, T., Kuboshima, M., Kuroiwa, N., Okazumi, S., Matsubara, H., Nomura, F., Takiguchi, M., and Hiwasa, T. (2004). Serological identification of trop2 by recombinant cdna expression cloning using sera of patients with esophageal squamous cell carcinoma. *International Journal of Cancer*, 112(6):1029–1035.
- [74] Neuenschwander, B., Branson, M., and Gsponer, T. (2008). Critical aspects of the bayesian approach to phase I cancer trials. *Statistics in Medicine*, 27(13):2420–2439.
- [75] Ni, I. B. P., Zakaria, Z., Muhammad, R., Abdullah, N., Ibrahim, N., Emran, N. A., Abdullah, N. H., and Hussain, S. N. A. S. (2010). Gene expression patterns distinguish breast carcinomas from normal breast tissues: the malaysian context. *Pathology-Research and Practice*, 206(4):223–228.
- [76] of Health, N. I. (2017). Nih’s definition of a clinical trial. <https://grants.nih.gov/policy/clinical-trials/definition.htm>.
- [77] O’Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, pages 33–48.
- [78] Paller, C. J., Bradbury, P. A., Ivy, S. P., Seymour, L., LoRusso, P. M., Baker, L., Rubinstein, L., Huang, E., Collyar, D., Groshen, S., et al. (2014). Design of phase I combination trials: recommendations of the clinical trial design task force of the nci investigational drug steering committee. *Clinical Cancer Research*, 20(16):4210–4217.
- [79] Pallmann, P., Wan, F., Mander, A. P., Wheeler, G. M., Yap, C., Clive, S., Hampson, L. V., and Jaki, T. (2019). Designing and evaluating dose-escalation studies made easy: The modest web app. *Clinical Trials*, page 1740774519890146.
- [80] Pan, H., Lin, R., Zhou, Y., and Yuan, Y. (2020). Keyboard design for phase I drug-combination trials. *Contemporary Clinical Trials*, page 105972.
- [81] Petit, C., Samson, A., Morita, S., Ursino, M., Guedj, J., Jullien, V., Comets, E., and Zohar, S. (2018). Unified approach for extrapolation and bridging of adult information in early-phase dose-finding paediatric studies. *Statistical methods in medical research*, 27(6):1860–1877.
- [82] Phan, J. (2018). Trial of stereotactic hypofractionated radioablative (HYDRA) treatment of laryngeal cancer. *ClinicalTrials.gov Identifier: NCT03114462*.

- [83] Reynolds, A. R. (2010). Potential relevance of bell-shaped and u-shaped dose-responses for the therapeutic targeting of angiogenesis in cancer. *Dose-response*, 8(3):dose-response.
- [84] Rogatko, A., Schoeneck, D., Jonas, W., Tighiouart, M., Khuri, F. R., and Porter, A. (2007). Translation of innovative designs into phase I trials. *Journal of Clinical Oncology*, 25(31):4982–4986.
- [85] Ruppert, A. S. and Shoben, A. B. (2018). Overall success rate of a safe and efficacious drug: Results using six phase 1 designs, each followed by standard phase 2 and 3 designs. *Contemporary Clinical Trials Communications*, 12:40–50.
- [86] Sambucini, V. (2008). A bayesian predictive two-stage design for phase II clinical trials. *Statistics in Medicine*, 27(8):1199–1224.
- [87] Shan, G. and Gerstenberger, S. (2017). Fisher’s exact approach for post hoc analysis of a chi-squared test. *PloS one*, 12(12):e0188709.
- [88] Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10(1):1–10.
- [89] Simon, R., Rubinstein, L., Arbut, S. G., Christian, M. C., Freidlin, B., and Collins, J. (1997). Accelerated titration designs for phase I clinical trials in oncology. *Journal of the National Cancer Institute*, 89(15):1138–1147.
- [90] Skolnik, J. M., Barrett, J. S., Jayaraman, B., Patel, D., and Adamson, P. C. (2008). Shortening the timeline of pediatric phase I trials: the rolling six design. *Journal of Clinical Oncology*, 26(2):190–195.
- [91] Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials*, 7(1):8–17.
- [92] Stylianou, M. and Follmann, D. A. (2004). The accelerated biased coin up-and-down design in phase I trials. *Journal of Biopharmaceutical Statistics*, 14(1):249–260.
- [93] Takeda, K., Taguri, M., and Morita, S. (2018). BOIN-ET: Bayesian optimal interval design for dose finding based on both efficacy and toxicity outcomes. *Pharmaceutical Statistics*, 17(4):383–395.
- [94] Tan, S.-B. and Machin, D. (2002). Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine*, 21(14):1991–2012.

- [95] Team, R. C. (2014). R: A language and environment for statistical computing <http://www.r-project.org>.
- [96] Thall, P. F. and Cook, J. D. (2004). Dose-finding based on efficacy–toxicity trade-offs. *Biometrics*, 60(3):684–693.
- [97] Thall, P. F. and Russell, K. E. (1998). A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*, pages 251–264.
- [98] Thall, P. F. and Simon, R. (1994). Practical bayesian guidelines for phase IIb clinical trials. *Biometrics*, pages 337–349.
- [99] Thall, P. F., Simon, R. M., and Estey, E. H. (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine*, 14(4):357–379.
- [100] Thorlund, K., Golchi, S., Haggstrom, J., and Mills, E. (2019). Highly efficient clinical trials simulator (HECT): Software application for planning and simulating platform adaptive trials. *Gates Open Research*, 3.
- [101] Wages, N. A., Conaway, M. R., and O’Quigley, J. (2011). Continual reassessment method for partial ordering. *Biometrics*, 67(4):1555–1563.
- [102] Wages, N. A. and Petroni, G. R. (2018). A web tool for designing and conducting phase I trials using the continual reassessment method. *BMC cancer*, 18(1):133.
- [103] Wang, Y.-G., Leung, D. H.-Y., Li, M., and Tan, S.-B. (2005). Bayesian designs with frequentist and bayesian error rate considerations. *Statistical Methods in Medical Research*, 14(5):445–456.
- [104] Wu, J. (2019). TG02 plus dose-dense or metronomic temozolomide followed by randomized phase II trial of TG02 plus temozolomide versus temozolomide alone in adults with recurrent anaplastic astrocytoma and glioblastoma. *ClinicalTrials.gov Identifier: NCT02942264*.
- [105] Yan, F., Mandrekar, S. J., and Yuan, Y. (2017a). Keyboard: a novel bayesian toxicity probability interval design for phase I clinical trials. *Clinical Cancer Research*, 23(15):3994–4003.
- [106] Yan, F., Mandrekar, S. J., and Yuan, Y. (2017b). Keyboard: a novel bayesian toxicity probability interval design for phase I clinical trials. *Clinical Cancer Research*, 23:3994–4003.

- [107] Yin, G., Chen, N., and Jack Lee, J. (2012). phase II trial design with bayesian adaptive randomization and predictive probability. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(2):219–235.
- [108] Yin, G., Chen, N., and Lee, J. J. (2018). Bayesian adaptive randomization and trial monitoring with predictive probability for time-to-event endpoint. *Statistics in biosciences*, 10(2):420–438.
- [109] Yin, G., Li, Y., and Ji, Y. (2006). Bayesian dose-finding in phase i/ii clinical trials using toxicity and efficacy odds ratios. *Biometrics*, 62(3):777–787.
- [110] Yin, G. and Lin, R. (2015). Comments on ‘competing designs for drug combination in phase I dose-finding clinical trials’ by m-k. riviére, f. dubois, and s. zohar. *Statistics in Medicine*, 34(1):13–17.
- [111] Yin, G. and Yuan, Y. (2009). Bayesian model averaging continual reassessment method in phase I clinical trials. *Journal of the American Statistical Association*, 104(487):954–968.
- [112] Yuan, Y., Hess, K. R., Hilsenbeck, S. G., and Gilbert, M. R. (2016a). Bayesian optimal interval design: a simple and well-performing design for phase I oncology trials. *Clinical Cancer Research*, 22(17):4291–4301.
- [113] Yuan, Y., Hess, K. R., Hilsenbeck, S. G., and Gilbert, M. R. (2016b). Bayesian optimal interval design: a simple and well-performing design for phase I oncology trials. *Clinical Cancer Research*, 22(17):4291–4301.
- [114] Yuan, Y., Lee, J. J., and Hilsenbeck, S. G. (2019a). Model-assisted designs for early-phase clinical trials: Simplicity meets superiority. *JCO Precision Oncology*, 3:1–12.
- [115] Yuan, Y., Lee, J. J., and Hilsenbeck, S. G. (2019b). Model-assisted designs for early-phase clinical trials: Simplicity meets superiority. *JCO Precision Oncology*, 3:1–12.
- [116] Yuan, Y., Lin, R., Li, D., Nie, L., and Warren, K. E. (2018a). Time-to-event bayesian optimal interval design to accelerate phase I trials. *Clinical Cancer Research*, 24(20):4921–4930.
- [117] Yuan, Y., Nguyen, H. Q., and Thall, P. F. (2017). *Bayesian designs for Phase I-II clinical trials*. Chapman and Hall/CRC.

- [118] Yuan, Y. and Yin, G. (2009). Bayesian dose finding by jointly modelling toxicity and efficacy as time-to-event outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(5):719–736.
- [119] Yuan, Y., Zhou, H., and Zhou, Y. (2018b). Phase I cancer clinical trial design: Single and combination agents. In *Biopharmaceutical Applied Statistics Symposium*, pages 205–233. Springer.
- [120] Zang, Y., Lee, J. J., and Yuan, Y. (2014). Adaptive designs for identifying optimal biological dose for molecularly targeted agents. *Clinical Trials*, 11(3):319–327.
- [121] Zhang, L. and Yuan, Y. (2016). A practical bayesian design to identify the maximum tolerated dose contour for drug combination trials. *Statistics in Medicine*, 35(27):4924–4936.
- [122] Zhou, H., Lee, J. J., and Yuan, Y. (2017). Bop2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. *Statistics in Medicine*, 36(21):3302–3314.
- [123] Zhou, H., Liu, F., Wu, C., Rubin, E. H., Giranda, V. L., and Chen, C. (2019a). Optimal two-stage designs for exploratory basket trials. *Contemporary Clinical Trials*, page 105807.
- [124] Zhou, H., Murray, T. A., Pan, H., and Yuan, Y. (2018a). Comparative review of novel model-assisted designs for phase I clinical trials. *Statistics in Medicine*, 37(14):2208–2222.
- [125] Zhou, H., Yuan, Y., and Nie, L. (2018b). Accuracy, safety, and reliability of novel phase I trial designs. *Clinical Cancer Research*, 24(18):4357–4364.
- [126] Zhou, Y., Lee, J. J., and Yuan, Y. (2019b). A utility-based bayesian optimal interval (U-BOIN) phase I/II design to identify the optimal biological dose for targeted and immune therapies. *Statistics in Medicine*, 38(28):S5299–S5316.
- [127] Zhou, Y., Lee, J. J., and Yuan, Y. (2020). Incorporating historical information to improve phase I clinical trial designs. *Unpublished*.
- [128] Zohar, S., Katsahian, S., and O’Quigley, J. (2011). An approach to meta-analysis of dose-finding studies. *Statistics in Medicine*, 30(17):2109–2116.

VITA

Yanhong Zhou was born in Qianxi, Guizhou, China. After completing her high school education at Shuixi Middle School, she entered Dongbei University of Finance and Economics and earned her Bachelor's degree in Hotel and Tourism Management in 2011. She received a Master's degree in Recreation parks and tourism management in August, 2014, and a Master's degree in Statistics in May, 2016 at West Virginia University. She has attended The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences since the August of 2016.