Novel Bayesian Adaptive Designs for Early Phase Oncology Clinical Trials

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Kristen M. Cunanan

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advised by Joseph S. Koopmeiners

August, 2015

Acknowledgments

I am extremely grateful for the research advisors and mentors that have helped me during my academic development. I express my deepest gratitude to my thesis advisor, Joe Koopmeiners, whose guidance, assistance, and unwavering support, especially during personal hardships, has helped me more than I think he will ever know.

I am very grateful to my committee members, Brad Carlin, Jim Hodges, and Jaime Modiano for taking the time to review and help improve my thesis.

Dedication

This thesis is dedicated to my loving family, but most especially, to my parents.

Abstract

Bayesian adaptive clinical trial designs are slowly gaining momentum in practice due to their accuracy, flexibility and efficiency in evaluating a novel drug. In this thesis, we propose novel Bayesian adaptive designs for early phase oncology trials. First, we discuss a Phase I-II trial design for therapeutic cancer vaccines and propose a two-stage approach for identifying the optimal vaccination schedule from multiple candidate vaccination schedules. We model binary outcomes for toxicity and immune response and a continuous outcome for the magnitude of immune response, conditional on a non-zero immune response. Our results suggest that incorporating more sources of information in a two-stage approach provides adequate power to identify the optimal schedule by trial completion. Next, we propose a novel Bayesian adaptive Phase I trial design that uses hierarchical modeling to share information across multiple patient populations, which may have different background standards-of-care. We propose hierarchical extensions for three models commonly used in Phase I clinical trials and propose three novel dose-finding guidelines that allow us to take full advantage of hierarchical modeling while protecting patient safety. We conclude by extending our hierarchical modeling approach to Phase I-II dose-escalation studies, where dose selection is based on both toxicity and efficacy. Our simulation results show that hierarchical modeling increases the probability of correctly identifying the maximum tolerated dose or optimal dose without increasing the rate of dose limiting toxicities. The results in this thesis are promising and motivate further research to investigate the practical challenges in implementing our proposed designs.

Contents

Acknowledgments **i**

Dedication **ii**

Abstract **iii**

List of Tables **vii**

List of Figures **xi**

1 Introduction **1**

1.1 Phase I Clinical Trials 2

1.2 Phase II Clinical Trials 4

1.3 Phase I-II Designs 5

1.4 Dissertation Objectives 7

2 A Bayesian Adaptive Phase I-II Trial Design for Therapeutic Cancer Vaccines **9**

2.1 Introduction 9

2.2 Probability Model and Trial Design 11

2.2.1 Probability Model and Prior Specification 11

2.2.2 Therapeutic Cancer Vaccine Trial Design 13

2.3 Simulation Study 17

2.3.1 Results 18

2.4 Extending to K Vaccine Therapies 22

2.5 Discussion 24
2.6 Supplementary Materials 27

3 Hierarchical Models for Sharing Information Across Populations in Phase I Dose-Escalation Studies 37
3.1 Introduction 37
3.1.1 Hierarchical Modeling in Phase I Oncology Trials 38
3.2 Dose-Toxicity Models 39
3.2.1 Power Model 39
3.2.2 Logistic Regression Model 40
3.2.3 Curve-Free Model 41
3.3 Dose-Finding Algorithm 42
3.3.1 Dose-Finding Guidelines 44
3.4 Simulation Study 51
3.4.1 Scenarios 52
3.4.2 Results 53
3.4.3 Exploring Other K 59
3.5 Discussion 61
3.6 Supplementary Materials 65

4 Efficacy/Toxicity Dose-Finding Using Hierarchical Modeling for Multiple Populations 75
4.1 Introduction 75
4.2 Models 77
4.2.1 Bivariate Binary Outcomes 78
4.2.2 Trinary Outcome 82
4.2.3 Hyperparameter Specification 84
4.3 Dose-Finding Algorithm 84
4.4 Simulation Study 88
4.4.1 Scenarios 89
4.4.2 Results 89
4.5 Discussion 95

5 Conclusion **97**

5.1 Summary of Major Findings 97

5.2 Future Work and Considerations 98

References **100**

List of Tables

2.1 Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2. Also presented are the probability of stopping the trial for an inconclusive result after stage 1 (Stop Trial), the expected overall sample size, EN $= 2 \times (n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2)$, and the average Stage 2 sample size for each schedule conditional on expanding to stage 2, $\bar{n}_2$. The correct outcome is in bold. 20

2.2 Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2. Also presented are the probability of stopping the trial for an inconclusive result after stage 1 (Stop Trial), the expected overall sample size, EN $= 2 \times (n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2)$, and the average Stage 2 sample size for each schedule conditional on expanding to stage 2, $\bar{n}_2$. The correct outcome is in bold. 21

2.3 Simulation results when $n_1 = 15$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2. Also presented are the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected overall sample size, EN $= 2 \times (n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2)$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold. 28

2.4 Simulation results when $n_1 = 20$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2. Also presented are the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected overall sample size, EN $= 2 \times (n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2)$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold. 29

2.5 Simulation results when $n_1 = 10$, $n = 50$. Changing: γ_T, γ_E, γ_θ and β (with the no. of stimulated stage 2's used to estimate n_2) is presented in the first column. Presented are: the probabilities of declaring a schedule better after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size/schedule, En= $n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2$, and the average Stage 2 sample size per schedule conditional on expanding to stage 2, $\bar{n}_2$. The correct outcome is in bold. 30

2.6 **Alt. Prior:** $\sigma \sim Uniform(0, 1)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold. 31

2.7 **Alt. Prior:** $\sigma \sim Uniform(0, 5)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold. 32

2.8 **Alt. Prior:** $\sigma \sim Uniform(0, 10)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr$(continuing onto stage 2) $\times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold. 33

2.9 **Alt. Prior:** $\mu \sim N(0, 1)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr$(continuing onto stage 2) $\times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold. 34

2.10 **Alt. Prior:** $\mu \sim N(0, 10)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr$(continuing onto stage 2) $\times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold. 35

2.11 **Alt. Prior:** $\mu \sim N(0, 100)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr$(continuing onto stage 2) $\times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold. 36

3.1 Results from 1000 simulated trials for the power model for different K. Below are the selection probabilities for the target dose. Results using HM are in bold, while results from an independent design are displayed in the next row. $K = 5$ group indices are listed for each Scenario within each K. 62

List of Figures

2.1 Outcome frequency for determining the best schedule when comparing four vaccination schedules. The white fill represents the outcome probabilities after the first stage; the grey fill represents the improved outcome probabilities after the second stage. In the "Inconcl." bar, the diagonal fill represents the percent of trials that stop after Stage 1 because the best schedule could not be determined using predictive probability with the maximum sample size enrolled. True toxicity and immune response rates (and magnitude) are displayed directly below each vaccination schedule. 25

3.1 **no DFG**: Display of dose-finding for a simulated trial when no DFGs are implemented in our dose-finding algorithm. White circles display the patient number; grey circles depict if the patient experienced a DLT. The true MTD_k is displayed at the top of the figure, directly below each population; the estimated MTD_k is displayed at the bottom of the figure for each population. 47

3.2 **1(m)**: Display of dose-finding for the same simulated trial when the "1(m)" DFGs are implemented in our dose-finding algorithm. White circles display the patient number; grey circles depict if the patient experienced a DLT; bolded circles represent if dose-escalation is allowed for any of the populations after patient enrollment. The true MTD_k is displayed at the top of the figure, directly below each population; the estimated MTD_k is displayed at the bottom of the figure for each population. . . . 48

3.3 **m/1(m+1)**: Display of dose-finding for the same simulated trial when the "m/1(m+1)" DFGs are implemented in our dose-finding algorithm. White circles display the patient number; grey circles depict if the patient experienced a DLT; bolded circles represent if dose-escalation is allowed for any of the populations after patient enrollment. The true MTD_k is displayed at the top of the figure, directly below each population; the estimated MTD_k is displayed at the bottom of the figure for each population. 49

3.4 **321**: Display of dose-finding for the same simulated trial when the "321" DFGs are implemented in our dose-finding algorithm. White circles display the patient number; grey circles depict if the patient experienced a DLT; bolded circles represent if dose-escalation is allowed for any of the populations after patient enrollment. The true MTD_k is displayed at the top of the figure, directly below each population; the estimated MTD_k is displayed at the bottom of the figure for each population. 50

3.5 Scenario 1 (top, left): all populations' dose-response curve are equivalent. Scenario 2 (top, right): all populations have the same MTD level, but slope increases with population index. Scenario 3 (middle, left): the MTD for each population is dispersed across all four dose-levels for the five populations. Scenario 4 (middle, right): the first three populations' MTD is dose level 1; the last two populations have MTD at dose level 2 and 3, respectively. Scenario 5 (bottom, left): population 1 terminates trial; the next two populations' MTD is dose level 1 the last two populations' MTD is dose level 2. Scenario 6 (bottom, left): similar to Scenario 4, except population 1 terminates trial. The MTD for each population is identified with an asterisk. 54

3.6 **Power Model**: Scenarios 1-3(left plots): Probability of correctly identifying the true MTD; Scenario 1-3 (right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis. 56

3.7 **Logistic Regression Model**: Scenarios 1-3(left plots): Probability of correctly identifying the true MTD; Scenario 1-3 (right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis. 58

3.8 **Curve-Free Model**: Scenarios 1-3(left plots): Probability of correctly identifying the true MTD; Scenario 1-3 (right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis. 60

3.9 **Power Model**: Scenarios 4-6(left plots): Probability of correctly identifying the true MTD; Scenario 4-6(right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis. 66

3.10 **Power Model**: Scenarios 1-3(left plots): Average number of patients treated at the true MTD; Scenario 1-3(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis. 67

3.11 **Power Model**: Scenarios 4-6(left plots): Average number of patients treated at the true MTD; Scenario 4-6(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis. 68

3.12 **Logistic Regression Model**: Scenarios 4-6(left plots): Probability of correctly identifying the true MTD; Scenario 4-6(right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis. 69

3.13 **Logistic Regression Model**: Scenarios 1-3(left plots): Average number of patients treated at the true MTD; Scenario 1-3(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis. 70

3.14 **Logistic Regression Model**: Scenarios 4-6(left plots): Average number of patients treated at the true MTD; Scenario 4-6(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis. 71

3.15 **Curve Free Model**: Scenarios 4-6(left plots): Probability of correctly identifying the true MTD; Scenario 4-6(right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis. 72

3.16 **Curve Free Model**: Scenarios 1-3(left plots): Average number of patients treated at the true MTD; Scenario 1-3(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis. 73

3.17 **Curve Free Model**: Scenarios 4-6(left plots): Average number of patients treated at the true MTD; Scenario 4-6(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis. 74

4.1 Thirteen combinations of dose-toxicity and dose-efficacy curves from Yin et al. (2006). The optimal dose based on different decision criteria are displayed above the pre-specified dose levels on the x-axis, and denoted as: (open circle) toxicity-efficacy odds ratio, (black dot) joint posterior probability of no toxicity with efficacy, (x) Zhang et al. (2006) decision rule with $\lambda = 0$, and (square) alteration: posterior probability of efficacy conditional on no toxicity, Zhang et al. (2006) decision rule with $\lambda = 0$. The dose-toxicity and dose-efficacy curves are represented with red and blue lines, respectively. The grey line displays the upper toxicity and lower efficacy limits for our posterior probabilities. **Scenario 1**: All five populations assume dose-response curves from Case 1. **Scenario 2**: Cases 3, 13, 13, 13, and 13 for populations 1, 2, 3, 4, and 5, respectively. **Scenario 3**: Cases 3, 9, 1, 8, 11. **Scenario 4**: Cases 6, 7, 8, 10, 13. . . 90

4.2 (Left column) Trial operating characteristics from 1000 simulated trials for Scenario 1 by population: (top) selection probability for the population-specific biologically optimal dose (BOD); (middle) percentage of dose-limited toxicities; (bottom) average number of patients treated at the population specific BOD, for the three hierarchical and independence designs. The first two bars (black and dark grey; labelled "PLR HM" and "PLR Ind.", respectively) present results using the parametric bivariate binary models. The next two bars (darker grey and grey; labelled "OR HM" and "OR Ind.", respectively) present results using the non-parametric bivariate binary models. The last two bars (light grey and white; labelled "Tri HM" and "Tri Ind.", respectively) present results for the parametric trinary model. For each scenario, the population's case-index used for data generation is presented on the x-axis. (Right column) Simulation results for Scenario 2. 92

4.3 (Left column) Trial operating characteristics from 1000 simulated trials for Scenario 3 by population: (top) selection probability for the population-specific biologically optimal dose (BOD); (middle) percentage of dose-limited toxicities; (bottom) average number of patients treated at the population specific BOD, for the three hierarchical and independence designs. The first two bars (black and dark grey; labelled "PLR HM" and "PLR Ind.", respectively) present results using the parametric bivariate binary models. The next two bars (darker grey and grey; labelled "OR HM" and "OR Ind.", respectively) present results using the non-parametric bivariate binary models. The last two bars (light grey and white; labelled "Tri HM" and "Tri Ind.", respectively) present results for the parametric trinary model. For each scenario, the population's case-index used for data generation is presented on the x-axis. (Right column) Simulation results for Scenario 4. 94

Chapter 1

Introduction

The clinical development of a novel drug is a long and expensive process. The new drug first goes through years of laboratory testing and sometimes animal testing before reaching clinical evaluation in humans. These pre-clinical studies investigate the drug's half-life, basic pharmacokinetics (i.e., what the body does to the drug) and pharmacodynamics (i.e., what the drug does to the body). The new drug then proceeds through three phases of clinical investigation before approval. The "early" and "middle" phase (Phase I and Phase II in the U.S., respectively) treat relatively small numbers of patients to evaluate safety and establish an optimal dose for further investigation in a large confirmatory trial (Phase III). Once approved, Phase IV post-market surveillance trials are often completed. These studies observe dynamic populations for side-effects related to the newly approved treatment.

In 2006, the United States Food and Drug Administration (FDA) released a *Critical Path Opportunities List* outlining 76 initial projects to guide investigators in an attempt to improve the drug development process. This document highlighted the need for innovative trial designs, especially those incorporating prior knowledge or accumulated information in the design. Subsequently, an increasing number of *adaptive* clinical trial designs have been proposed and implemented due to their ability to flexibly and efficiently evaluate novel treatments (Chow et al., 2008). In February 2010, the FDA circulated a non-specific draft guidance on adaptive clinical trials, which was generally supportive of properly employed adaptive designs (Food and Drug Administration et al.,

2010).

Broadly speaking, an adaptive clinical trial design refers to a design that allows modifications to the trial and/or statistical procedure while the trial is ongoing (Chow et al., 2005). The Pharmaceutical Research Manufacturers Association Working Group on Adaptive Designs makes a further distinction, stating that any design that allows modifications based on accumulating data is considered adaptive (Gallo et al., 2006). Chow et al. (2008) and Chow and Corey (2011) discuss the advantages and limitations of commonly used adaptive clinical trial designs. Adaptive designs can be either outcome or covariate (patient characteristics) dependent. Adaptive designs have been proposed that alter a number of trial characteristics based on accumulating data, including but not limited to dosage, sample size and randomization allocation. These designs allow investigators to efficiently evaluate several characteristics of a novel treatment in a single trial. This dissertation will investigate novel Bayesian adaptive methods for early phase cancer clinical trials.

1.1 Phase I Clinical Trials

The primary objective of a Phase I clinical trial is to evaluate the safety profile of a novel treatment and identify the maximum tolerated dose (MTD), defined as the highest dose with an acceptable toxicity rate less than some pre-defined target toxicity level (TTL, typically 0.3 in Phase I cancer trials). The primary endpoint in Phase I clinical trials is typically a binary indicator of whether or not the patient experienced a dose-limiting toxicity (DLT). A DLT is an adverse event, or toxic side-effect, that is severe enough to prevent increasing the dosage. In Phase I oncology trials, investigators define adverse events and classify their severity using the Common Terminology Criteria for Adverse Events (CTCAE). Typically, a subset of the Grade 3 and 4 toxicities outlined in the CTCAE are used to define a DLT. A second objective of Phase I clinical trials is to identify the appropriate dose for future study in Phase II clinical trials. Historically, researchers assumed both the probability of toxicity and the probability of tumor response increased monotonically with dose and the MTD was considered the dose most likely to result in tumor response. However, this monotonicity assumption may not be valid

for newer biologically targeted treatments, which motivates designs that consider both toxicity and efficacy during dose-finding. Finally, due to the severe toxicity of novel cancer treatments, Phase I cancer trials enroll patients for whom standard treatments have failed. As a result, there is a desire to treat as many patients as possible at doses close to the MTD to maximize the number of patients receiving a biologically active dose, which creates additional challenges for Phase I clinical trials in oncology.

Phase I clinical trials take the form of dose-escalation studies, where initial patients are treated at the lowest dose-level and subsequent patients are treated at progressively higher dose-levels until the MTD is identified. Phase I dose-escalation studies can broadly be classified as rule-based or model-based. Rule-based designs are simple and easy to implement. Dose escalation or de-escalation is based on the presence or absence of a DLT in the previous cohort. The most commonly used rule-based design is the $3+3$ design (see e.g. Storer (1989)). The traditional $3+3$ design escalates in cohorts of three subjects until the MTD is identified. In the $3+3$, the MTD is defined as the highest dose with at least six patients treated and no more than one patient experiencing a DLT. Other examples of rule-based designs include the biased coin design (Stylianou and Flournoy, 2002), which escalates with probability $\frac{TTL}{1-TTL}$, where TTL is the target DLT rate, if the previous cohort did not experience toxicity and always de-escalates if the previous cohort experienced a toxicity, and the broad class of up-and-down designs proposed to improve upon the $3+3$ and biased-coin design (Ivanova et al., 2003). Currently, the $3+3$ design is the most commonly used dose-finding design in the United States (Riviere et al., 2015) even though several authors have shown that the $3+3$ will reach the MTD slowly, treat a small number of patients at the MTD, and is likely to produce an MTD that is futile when studied in Phase II trials (Garrett-Mayer, 2006; Goodman et al., 1995; Iasonos et al., 2008; O'Quigley and Shen, 1996; Jaki et al., 2013; Braun, 2014).

Alternatively, dose-escalation studies can be model-based, where dose-finding is guided by a formal statistical model for the dose-toxicity relationship. Unlike the 3+3 design, escalation or de-escalation in a model-based design is based on all available data, rather than only considering the outcomes for the previous cohort. The continual reassessment

method (CRM) (O'Quigley et al., 1990) is the most commonly used model-based design. The basic CRM uses a simple parametric model for the dose-toxicity relationship and is typically implemented under the Bayesian statistical paradigm. The dose-finding algorithm updates the dose-toxicity curve after each cohort, with the next cohort treated at the current estimate of the MTD based on all available data. Additional modifications to the CRM, such as starting at the lowest dose level and not allowing dose-levels to be skipped during escalation, have been proposed to ensure patient safety (Goodman et al., 1995; Korn et al., 1994). Numerous extensions to the CRM have been proposed to improve estimation and accommodate the challenges faced by clinical researchers (Babb et al., 1998; Cheung and Chappell, 2000; Legedza and Ibrahim, 2001; Neuenschwander et al., 2008; Yin and Yuan, 2009; Liu et al., 2015; O'Quigley and Paoletti, 2003; Braun and Wang, 2010; Iasonos and O'Quigley, 2012, 2013).

1.2 Phase II Clinical Trials

Once a new treatment has been shown to be safe in Phase I, the next step in the drug development process is to evaluate its potential efficacy in a Phase II clinical trial. Phase II is an important step in the drug development process that is often sub-divided into Phase IIA and Phase IIB. Phase IIA trials typically consist of single-arm trials where measures of treatment efficacy are compared to historical benchmarks of performance, while Phase IIB trials are typically small randomized trials to compare the new treatment to the standard-of-care or other experimental treatments. In oncology trials, the primary endpoint for Phase IIA trials is typically a binary measure of tumor response (i.e., did the tumor shrink or not), while Phase IIB trials often consider survival endpoints like time-to-disease progression.

The most commonly used Phase IIA design in oncology is Simon's two-stage design (Simon, 1989). Simon's two-stage design is a one-arm trial that compares the response rate of the new drug to historical standards. In Stage 1, initial subjects are enrolled and the trial terminates for futility if there is substantial evidence against the alternative hypothesis. Otherwise, the trial continues to Stage 2 where the observed response rate is compared to the historical response rate for the standard-of-care, assumed to be fixed

and known. Simon (1989) discusses both optimal and minimax two-stage designs, which minimize the expected or maximum sample size, respectively. Alternately, continuous monitoring can be implemented in one-arm Phase IIA designs using Bayesian predictive probabilities (Lee and Liu, 2008). Continuous monitoring can reduce the expected sample size compared to a simpler two-stage design, but comes at the cost of additional complexity. In general, adaptation in the context of Phase IIA trials focuses primarily on sequential testing, and involves a trade-off between the additional complexity of more frequent monitoring and the benefits of reducing the time and number of patients needed to achieve a significant result.

Phase IIB trials are randomized trials comparing the new treatment to the standard-of-care or other experimental treatments. Standard approaches to randomized, multi-arm clinical trials are well-understood and readily accepted. The challenge to completing these trials in oncology is a result of the need to rigorously evaluate the new treatments with limited resources (typically, only 40 - 200 patients are treated in Phase IIB trials), combined with the desire to provide patients with access to the best treatment available for their condition. This motivates investigators to consider more complex, adaptive approaches to maximize resources. The earliest adaptive design for Phase IIB trials is the "play-the-winner" rule (Zelen, 1969). Under the this rule, the next patient will be treated with the same treatment as the current patient if the current patient responds, but will be treated with the alternate treatment if the current patient does not respond. Adaptive randomization, which alters the randomization ratio based on accumulating data from the trial, has been proposed as an approach to maximize the number of patients receiving the optimal treatment in a clinical trial (Thall and Wathen, 2007; Berry et al., 2010). The utility of adaptive randomization is currently a topic of debate in the literature (Korn and Freidlin, 2011; Lee et al., 2012; Thall et al., 2015).

1.3 Phase I-II Designs

Historically, Phases I and II have been treated as two distinct stages in the development of a new cancer treatment. In Phase I, the MTD is identified under the assumption that the probabilities of DLT and tumor response increase monotonically with dose and, as

a result, the MTD is the dose most likely to achieve an efficacious response with acceptable toxicity. Then, in Phase IIA, the efficacy of the MTD is evaluated to determine if the new treatment should advance in the development process. An alternate approach, referred to as Phase I-II designs, is to evaluate both safety and efficacy in a single trial that considers the trade-off between efficacy and toxicity during dose-finding. These designs are appealing because they can detect a situation where the drug's efficacy may plateau or diminish after a certain dose level but toxicity continues to increase, in which case further escalation is undesirable, or where the MTD has unacceptable efficacy, in which case the treatment is unlikely to achieve a successful outcome in Phase II.

There are two major components to a Phase I-II trial design: the joint probability model for efficacy and toxicity, and the metric for evaluating the trade-off between efficacy and toxicity. Several authors have used *copula models* to specify the joint probability of efficacy and toxicity (Thall and Cook, 2004; Braun, 2002), while a non-parametric approach (Yin et al., 2006) has also been considered. However, these models incorporate complex correlation structures for efficacy and toxicity that may be difficult to estimate and, in some cases, simulation results have illustrated that a simple model that ignores the correlation between efficacy and toxicity may be adequate (Cunanan and Koopmeiners, 2014). In addition, several approaches have been proposed for evaluating the trade-off between efficacy and toxicity, including: efficacy/toxicity trade-off contours (Thall and Cook, 2004), a weighted Euclidean distance from target probabilities of efficacy and toxicity (Braun, 2002) and efficacy-toxicity odds ratios (Yin et al., 2006). Alternatively, many authors have proposed combining the two outcomes into one outcome (Thall and Russell, 1998; Zhang et al., 2006), to evaluate the trade-off between efficacy and toxicity without drastically increasing the parameter space. Once the model and trade-off metric are specified, these designs proceed similarly to the CRM, with each cohort treated at the current estimate of the optimal dose based on the data from all previous cohorts. Numerous authors have expanded on the basic idea of a Phase I-II design that considers both efficacy and toxicity for a number of specialized scenarios (Thall and Russell, 1998; Braun, 2002; Zohar and Chevret, 2007; Thall and Cook, 2004; Zhang et al., 2006; Yin et al., 2006; Zhong et al., 2012; Koopmeiners and Modiano, 2014).

1.4 Dissertation Objectives

In this dissertation, we propose novel Bayesian adaptive trial designs for early phase cancer clinical trials. Phase I-II designs have been developed primarily in the context of dose-escalation studies. However, there are some clinical applications where dose-escalation is not of primary interest, such as therapeutic cancer vaccines. In Chapter 2, we discuss a Bayesian adaptive Phase I-II clinical trial design for a novel autologous vaccine for high grade meningiomas. In this application, investigators are interested in comparing different vaccination schedules, rather than different dose-levels. We develop a two-stage design using binary toxicity and immune response outcomes and a continuous outcome for the magnitude of immune response to determine the optimal vaccination schedule. If a conclusive result cannot be reached after the first stage, Bayesian predictive probabilities are used to determine the sample size needed in Stage 2 to achieve a conclusive result.

In Chapter 3, we investigate hierarchical modeling in the context of Phase I dose-escalation studies. Researchers are often interested in evaluating the safety of a new drug in multiple patient populations with different background standards-of-care. In this case, researchers typically complete independent Phase I trials in the different populations, which may or may not result in different estimates of the MTD across populations. This is an expensive and time-consuming process that ignores the fact that information regarding the MTD in one population can also be found in the outcomes of patients in other populations. We consider a novel trial design that uses hierarchical modeling to share information across populations studied in multiple, parallel Phase I clinical trials. We propose hierarchical extensions of three commonly used dose-response models for Phase I trial designs, and propose dose-finding guidelines that allow us to take full advantage of hierarchical modeling while protecting patient safety. Our simulation results indicate hierarchical modeling increases the probability of correctly identifying the MTD and the average number of patients treated at the MTD, without increasing the rate of DLTs.

In Chapter 4, we extend this work to evaluate the use of hierarchical modeling in Phase

I-II dose-finding trials that consider efficacy in addition to toxicity. We propose hierarchical extensions of three commonly used dose-response models for multiple outcomes. Similar to Chapter 3, our results suggest hierarchical modeling increases the probability of correctly identifying the optimal dose and the average number of patients treated at the optimal dose for each population, with a minimal increase in the toxicity rate for some cases. Finally, in Chapter 5 we summarize our findings and discuss directions for future research in this clinically important area.

Chapter 2

A Bayesian Adaptive Phase I-II Trial Design for Therapeutic Cancer Vaccines

2.1 Introduction

The traditional approach to Phase I clinical trials in cancer (i.e. dose-escalation studies to identify the MTD) was developed primarily to evaluate the safety of new chemotherapeutic agents. Recently, the field of cancer therapeutics has expanded to include other approaches to cancer treatment including therapeutic cancer vaccines, which boost the immune system's ability to treat an existing cancer by using weakened or killed cancer cells in conjunction with a specific cancer-associated antigen or modified immune cells that express such an antigen (National Cancer Institute, 2006). Therapeutic cancer vaccines can either be autologous or allogenic, depending on the origin of the cancer cells. Autologous vaccines use cancer cells from the individual patient, while allogenic vaccines use cells from another patient (National Cancer Institute, 2006). A large number of cancer-associated antigens have been identified. Furthermore, researchers often add adjuvants to further stimulate the immune system. Hence, a vast number of potentially efficacious vaccine therapies, from various combinations of antigens and adjuvants, require investigation. Ideally, we want to identify the combination that yields not only

the largest response rate and magnitude, but minimizes the rate of toxic side effects.

Researchers at the University of Minnesota are interested in completing a Phase I-II clinical trial of a novel autologous vaccine for high grade meningiomas, an aggressive form of brain tumor. Toxicity will be measured by the frequency of dose limiting toxicities (DLTs), while efficacy will be evaluated by the presence and magnitude of immune response. In this case, our efficacy outcome is a combination of a binary indicator for the presence of immune response and, conditional on observing a response, a continuous measure of the magnitude of response. The researchers are not interested in completing a dose-escalation study. In fact, they are not interested in comparing multiple dose levels at all, but are rather interested in comparing two different vaccination schedules, which will deliver the same dose level with differing frequencies. In the first vaccination scheme, patients receive the regimen daily for 4 days every 4 weeks for the first 12 weeks, then once every 4 weeks until disease progression or 1 year. In the second schedule, vaccination is once every week for the first 12 weeks, then once every 4 weeks until disease progression or 1 year.

The literature investigating dose-schedule optimization in the context of Phase I dose-escalation studies is limited (Thall et al., 2013; Braun et al., 2005; Zhang and Braun, 2013). In most cases, these designs assume the competing dose-schedules are nested (Braun et al., 2005; Zhang and Braun, 2013), which implies a natural ordering of the schedules. In contrast, the schedules in our motivating example are not nested and have no natural ordering. Recently Thall et al. (2013) investigated simultaneous optimization of dose and schedule with non-nested schedules in the context of a Phase I-II clinical trial. Their design evaluated efficacy by considering the time-to-tumor response. In contrast, we propose to evaluate the efficacy of therapeutic cancer vaccines using the presence and magnitude of immune response. Furthermore, we will take advantage of the lack of an ordering for our vaccination schedules by randomizing patients to each vaccination schedule, rather than using a deterministic dose- or schedule-finding algorithm for assigning patients to a schedule. Numerous randomized, multistage Phase II designs that evaluate several experimental treatments have been proposed (Strauss and Simon, 1995; Yao and Venkatraman, 1998; Yao et al., 1996; Thall et al., 1989), but these

designs assume Phase II objectives and thus require large sample sizes after the toxicity profile has already been established in previous trials.

We propose a two-stage randomized trial design for selecting the best vaccination schedule from several candidate schedules, incorporating objectives from both Phase I and Phase II trials. In the case of two candidate vaccination schedules, we first randomize patients to one of the two schedules in Stage 1. At the end of Stage 1, we compare the rate of DLTs and immune response for each schedule to pre-specified minimum performance standards. Schedules that meet the minimum performance standard are then compared to each other using the magnitude of immune response, to identify the better schedule. If the superiority of a single schedule cannot be established after the first stage, Bayesian predictive probabilities are used to determine the additional sample size required to identify the better vaccination schedule in the second stage.

The remainder of this chapter proceeds as follows. In Section 2.2, we describe probability models for DLTs, the presence and magnitude of immune response and propose a two-stage design for the simple case of comparing two vaccination schedules. In Section 2.3, we present simulation results evaluating the operating characteristics of our proposed design and discuss its overall performance. In Section 2.4, we describe extending our design to K schedules and present initial simulation results evaluating its performance. Section 2.5 concludes this manuscript with a brief discussion of the benefits and practicality of our proposed design.

2.2 Probability Model and Trial Design

2.2.1 Probability Model and Prior Specification

We begin by introducing notation and presenting a joint probability model for toxicity and immune response. For now, we suppress the subscripts i, which indicates patient, and k, which indicates vaccination schedule, to simplify notation. Let X be a binary indicator of whether the subject experienced DLT, which follows a Bernoulli distribution with probability π_T. Immune response has two components: a binary indicator of

whether the subject had an immune response, and a continuous measure of the magnitude of response for subjects experiencing an immune response. Let Y be a Bernoulli random variable with probability π_E that takes the value 1 if a subject had an immune response and 0 otherwise, and let Z be the continuous measure of the magnitude of response, which follows a conditional distribution with $Z|Y = 1 \sim logNormal(\mu, \sigma^2)$ and $f(Z|Y = 0) = 0$ with probability 1. Note that the conditional probability density function $f(Z|Y = 0)$ has a point mass at zero so the marginal distribution of Z is not absolutely continuous. It was recently shown that the correlation between efficacy and toxicity could not be estimated reliably given the limited sample sizes found in Phase I-II clinical trials but also that failing to model this correlation had little impact on the operating characteristics of the trial (Cunanan and Koopmeiners, 2014). We therefore assume independence between DLTs and immune response to avoid unnecessarily increasing the complexity of our model.

Suppose that n patients are randomized to each vaccination schedule and we observe the following 3-tuples for each patient, where the full dataset for schedule k is $\mathbf{d_k} = \{(x_{k,1}, y_{k,1}, z_{k,1}), (x_{k,2}, y_{k,2}, z_2), \ldots, (x_{k,n}, y_{k,n}, z_{k,n})\}$. Then the joint likelihood for immune response and toxicity is:

$$L(\pi_{\mathbf{T}}, \pi_{\mathbf{E}}, \mu, \sigma^{\mathbf{2}}|\mathbf{d}) \propto \prod_{k=1}^{K} \prod_{i=1}^{n} \pi_{T,k}^{x_{k,i}} (1-\pi_{T,k})^{1-x_{k,i}} \left(\pi_{E,k} \times f(z_{k,i}|\mu_k, \sigma_k^2)\right)^{y_{k,i}} (1-\pi_{E,k})^{1-y_{k,i}},$$

where $f(\cdot)$ is a log Normal density function.

We must specify a prior distribution for all parameters to complete our Bayesian analysis. We specify $Beta(1, 1)$ priors for both the toxicity and immune response rates, π_T and π_E, respectively. This prior specification is identical to a $Uniform(0, 1)$; it has a conjugate distributional form and is thus computationally preferred. We set $\mu \sim Normal(0, \tau^2 = 3)$ for the average magnitude of non-zero immune responses (on the log scale). Lastly, we use a $Uniform(0, b = 2)$ prior on σ, specifying an upper bound. This is the Jeffrey's prior for σ with known μ. We originally considered the conjugate inverse gamma prior for σ^2, but its performance was poor. We will complete a sensitivity analysis, varying τ^2 and b, to evaluate the sensitivity of the results found in Section 2.3 to the priors for μ and σ.

The joint posterior distribution, $p(\pi_E, \mu, \sigma^2|\mathbf{d})$, is not available in closed form. Instead, we approximate the posterior distribution by sampling from the posteriors $p(\pi_E|\mathbf{d})$ and $p(\mu, \sigma^2|\mathbf{d})$ using the Gibbs sampler. In zero-inflated models, a "data augmentation" step is incorporated into the Gibbs sampler to approximate the posterior distributions (Ghosh et al., 2006). Summaries of the posterior (i.e., posterior mean, median, etc.) are calculated by summarizing the Gibbs draws from the posterior.

2.2.2 Therapeutic Cancer Vaccine Trial Design

We now present a randomized, two-stage study design for evaluating the safety and immune response of therapeutic cancer vaccines. Our basic approach is to randomize patients between the two vaccination schedules, determine if the probabilities of a DLT and immune-response satisfy a minimum level of clinical performance, then if necessary, select the schedule with the largest average magnitude of immune response (defined as a function of the immune response rate, as well). If Stage 1 does not give a conclusive result, we continue to Stage 2 and use predictive probabilities to determine the sample size required for Stage 2. We first consider the simple case of two vaccination schedules, as in our motivating example, and discuss generalizing to K schedules in Section 2.4.

In Stage 1, we randomize n_1 patients to each vaccination schedule and determine which, if any, of the schedules achieves a minimum performance level in response rates. Let $\overline{\pi}_T$ be the pre-specified maximum acceptable DLT rate and $\underline{\pi}_E$ be the pre-specified minimum acceptable immune response rate. We define a schedule, k to be acceptable if the posterior probabilities of the two events $\pi_{T,k} < \overline{\pi}_T$ and $\pi_{E,k} > \underline{\pi}_E$ exceed pre-specified thresholds γ_T and γ_E, respectively, i.e.,

$$\begin{aligned} Pr(\pi_{T,k} < \overline{\pi}_T|\mathbf{d_k}) &> \gamma_T \\ \text{and } Pr(\pi_{E,} > \underline{\pi}_E|\mathbf{d_k}) &> \gamma_E, \end{aligned} \tag{2.1}$$

similar to the admissibility criteria of Thall and Russell (1998) and Thall and Cook (2004). However, we use more stringent response requirements for the associated posterior distributions. In traditional dose-finding studies, investigators understand the practical need to treat patients below and above the MTD, especially early in the trial.

A primary objective in designing a Phase I dose finding algorithm is to maximize the number of patients treated at the true MTD. Traditionally, γ_E and γ_T are low (between 0.05 and 0.20, Berry et al. (2010)) since we are making a decision with often only 3, 6, or 9 patients, early in the trial. We use similar criteria to Thall and Russell (1998) and Thall and Cook (2004), but require a smaller type 1 error, so to speak, in making a decision after stage 1 using 10 or 15 patients, given the larger sample size compared to dose-finding cohorts.

After Stage 1, there are four possible outcomes, as follows. If both schedules do not satisfy Criteria (2.1) with n_1 patients treated in each schedule arm, we stop the trial and declare both vaccination schedules to be clinically unacceptable. If only one schedule is acceptable, we stop the trial and declare the acceptable schedule to be the better schedule. Alternately, if both satisfy Criteria (2.1), we compare the magnitudes of immune response and potentially expand the trial as described below.

Define θ_k to be the expected magnitude of immune response for schedule k, E(Z_k), i.e.,

$$\begin{aligned}\theta_k &= P\left(Z_k > 0|k\right) E\left(Z_k|Z_k > 0, k\right) \\ &= \pi_{E,k} exp\left(\mu_k + \frac{1}{2}\sigma_k^2\right) \\ &= exp\left(\mu_k + \frac{1}{2}\sigma_k^2 + log\pi_{E,k}\right). \end{aligned} \tag{2.2}$$

We can calculate θ_k for each iteration of the Gibbs sampler, and use these samples to approximate the posterior for θ_k. To determine superiority when both schedules are acceptable, we compare the two posterior expected magnitudes of immune response. *A priori* we assume equal variances for the two schedules, i.e., $\sigma_k^2 = \sigma^2$ for $k = 1, 2$, which allows cancellation of σ^2 when considering the ratio of the two schedules' expected magnitudes of immune response, θ_2/θ_1. This is reasonable since Phase I-II clinical trial sample sizes are limited and consequently sparse.

We use the posterior probability that $\theta_2 > \theta_1$ to determine whether a schedule has

a superior expected magnitude of immune response. If $Pr\left(\theta_2 > \theta_1|\mathbf{d}\right) > 1-\gamma_\theta$, we declare Schedule 2 to be the better schedule. If $Pr\left(\theta_2 > \theta_1|\mathbf{d}\right) < \gamma_\theta$, we declare Schedule 1 to be better. This is equivalent to comparing the $(1-2\gamma_\theta)\%$ Bayesian credible interval (CI) to 1 and declaring superiority if the CI is either completely above or completely below 1. If superiority cannot be achieved, we use posterior predictive probabilities to calculate the additional sample size required to achieve a definitive result.

Posterior predictive probabilities (PP) provide the probability of observing a conclusive result after Stage 2, conditional on the data observed in stage 1 and the stage-two sample size for each schedule, n_2. Our goal is to identify $2 * n_2$ such that the predictive probability of reaching a conclusive result exceeds β. We define some notation to simplify our expression for PP. Let $A_k = I\{Pr(\pi_{T,k} < \overline{\pi}_T|\mathbf{d_k}) > \gamma_T\} \times I\{Pr(\pi_{E,k} > \underline{\pi}_E|\mathbf{d_k}) > \gamma_E\}$ for $k = 1, 2$. For each $n_2 = 1, 2, 3, \ldots, n-n_1$, we can calculate the predictive probabilities as

$$\begin{aligned} PP &= E_{\mathbf{d}_*}(\{A_1 = 1 \cap [A_2 = 0 \cup Pr[\theta_2 > \theta_1|\mathbf{d}, \mathbf{d}^*] < \gamma_\theta)\} \\ &\quad \cup \{A_2 = 1 \cap [A_1 = 0 \cup Pr[\theta_2 > \theta_1|\mathbf{d}, \mathbf{d}^*] > 1-\gamma_\theta)\}) \\ &= \sum_{\mathbf{d}^*} Pr(\mathbf{d}^*|\mathbf{d}) \times (A_1 \times I\{A_2 = 0 \cup Pr[\theta_2 > \theta_1|\mathbf{d}, \mathbf{d}^*] < \gamma_\theta)\} \\ &\quad + A_2 \times I\{A_1 = 0 \cup Pr[\theta_2 > \theta_1|\mathbf{d}, \mathbf{d}^*] > 1-\gamma_\theta)\}), \end{aligned} \tag{2.3}$$

where $\mathbf{d}^*$ is the unobserved set of Stage 2 patients' outcomes. Note we can generate $\mathbf{d}^*$ using the assumed models for (X, Y, Z) and sample $(\pi_T, \pi_E, \mu, \sigma^2)$ from the joint posterior distributions. We compute (2.3) for $n_2 = 1, 2, \ldots$ until $PP \geq \beta$ or $n_2 = n - n_1$, where n is a pre-specified maximum sample size for each schedule (Berry et al., 2010). If the predictive probability exceeds β for some $n_2 = 1, 2, \ldots, n - n_1$, we randomize n_2 subjects to each schedule and compare the schedules' mean magnitudes of immune response, as described above.

In summary, our proposed design proceeds as follows:

- **Stage 1**:

    1. Randomize $2 * n_1$ patients in a 1:1 ratio to the two vaccination schedules

2. Evaluate $Pr(\pi_T < \overline{\pi}_T|\mathbf{d}) > \gamma_T$ and $Pr(\pi_E > \underline{\pi}_E|\mathbf{d}) > \gamma_E$ for both schedules to determine if they meet the minimum performance standard. There are four possible outcomes:
   - Neither schedule is acceptable: terminate the study and declare both schedules ineffective.
   - Schedule 1 is acceptable and Schedule 2 is not: terminate the study and declare Schedule 1 the better schedule.
   - Schedule 2 is acceptable and Schedule 1 is not: terminate the study and declare Schedule 2 the better schedule.
   - Both schedules are acceptable: compare the schedules by the magnitude of immune response.
3. Compare the magnitude of immune response by considering $Pr\left(\theta_2 > \theta_1|\mathbf{d}\right)$. There are three possible outcomes:
   - $Pr\left(\theta_2 > \theta_1|\mathbf{d}\right) > 1 - \gamma_\theta$: declare Schedule 2 the better schedule.
   - $Pr\left(\theta_2 > \theta_1|\mathbf{d}\right) < \gamma_\theta$: declare Schedule 1 the better schedule.
   - $\gamma_\theta < Pr\left(\theta_2 > \theta_1|\mathbf{d}\right) < (1 - \gamma_\theta)$: determine the sample size required to achieve a conclusive result using predictive probabilities.
4. Calculate the sample size required to achieve a conclusive result in Stage 2, $2 * n_2$, using (2.3): There are two possible outcomes:
   - $n_2 > n - n_1$, where n is a pre-specified maximum sample size for each schedule: terminate the study and declare an equivocal result.
   - $n_2 \leq n - n_1$: continue to Stage 2.

- **Stage 2**:

1. Randomize $2 * n_2$ patients in a 1:1 ratio to the two vaccination schedules.
2. Compare the two schedules and identify the better schedule following steps 2 and 3 in Stage 1. After Stage 2, the possible outcome are:
   - Schedule 1 is the better schedule.
   - Schedule 2 is the better schedule.
   - Both schedules are acceptable but we cannot identify an better schedule.

2.3 Simulation Study

We conducted a simulation study to evaluate the operating characteristics of our proposed design. Following our motivating example, we consider two vaccination schedules, which we will refer to as Schedules 1 and 2. Recall that in our design, we define a vaccination schedule as minimally clinically viable for both toxicity and immune response if Criteria (2.1) is satisfied. For this simulation, we require at least a 50% immune response rate, $\underline{\pi}_E = 0.5$, and no more than a 25% toxicity rate, $\overline{\pi}_T = 0.25$. These are typical immune response and toxicity rates in oncology trials. This means a schedule being acceptable if $\pi_T < 0.25$ and $\pi_E > 0.5$, and a schedule is unacceptable if $\pi_T \geq 0.25$ or $\pi_E \leq 0.5$. To determine the operating characteristics of our design, we consider the following scenarios: both schedules are unacceptable because the toxicity rate, immune response rate, or both are unacceptable; one schedule is acceptable and the other schedule is not; and finally, both schedules are acceptable but one schedule is better because it has a higher response rate or magnitude of immune response.

A number of design parameters must be specified to implement our design: n_1, n, γ_T, γ_E, γ_θ and β. These design parameters can be tuned via simulation to the achieve the desired operating characteristics of the study. We completed numerous small simulations to identify the optimal combination of design parameters for our example and have chosen the following combination of design parameters: $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Simulation results assuming these design parameters are shown in the next section. Additional simulation results evaluating the impact of varying these design parameters will be discussed and are presented in the Supplementary Materials in Table 2.5.

All simulations were completed in R version 3.1.1 and Gibbs sampling was completed in JAGS 3.1.0 as called from R using *rjags* (Plummer, 2011). 1000 simulated trials were completed for each scenario. Within each trial, 5000 iterations were kept for inference following a period of 5000 iterations for burn-in. If a trial is found inconclusive after stage 1, we simulate 100 Stage 2 sub-trials to calculate the predictive probability and additional $2 * n_2$ required to determine superiority.

2.3.1 Results

Table 2.1 presents results for our null scenarios. In the first case, we assume both schedules have an unacceptable toxicity rate of 25% and immune response rate of 50%, i.e. ($\pi_T = 0.25$, $\pi_E = 0.5$). Next, we consider the case where both schedules have acceptable immune response rates but are unsafe, i.e., (0.25, 0.9). Finally, we consider the case where both are safe but have unacceptable immune response rates, i.e., (0.05, 0.5). We will primarily evaluate our design by considering the probability of reaching each of the various conclusions after Stages 1 and 2. After Stage 1, the four conclusions are: neither schedule is acceptable (none), Schedule 1 is better (Schedule 1), Schedule 2 is better (Schedule 2) and both are acceptable but we are unable to identify a better schedule (Inconcl.). The same conclusions can be reached after Stage 2. In addition, we also present the probability that the trial was inconclusive but did not continue onto stage 2 because the sample size required to achieve the desired predictive probability exceeded our pre-specified maximum (Stop Trial). The probabilities of reaching the various conclusions after each stage sum to 1. In addition, we also present the expected overall sample size, EN $= 2 \times (n_1 + Pr(\text{continuing onto Stage 2}) \times \bar{n}_2)$, where $\bar{n}_2$ is the average sample size per schedule in stage 2 conditional on expanding to stage 2.

From Table 2.1, we see that our design has acceptable performance in correctly identifying both schedules as unacceptable after stage 1. For ($\pi_T = 0.25$, $\pi_E = 0.5$), 79% of all simulated trials correctly identify both schedules as unsafe and futile, while 20% of all simulated trials incorrectly select either Schedule 1 or Schedule 2. The expected overall sample size is 20.2 patients, suggesting that our design is unlikely to continue to stage 2. When both schedules have high immune response rates, i.e., $\pi_E = 0.9$, but are unsafe, i.e., $\pi_T = 0.25$, our design correctly concludes that neither schedule is acceptable in 61% of all simulated trials. Alternatively, when both schedules are safe but futile, i.e., ($\pi_T = 0.05$, $\pi_E = 0.5$), our design correctly identifies neither schedule as acceptable in 38% of all simulated trials. This is due to our choosing a lower critical threshold value for declaring efficacy because efficacy will be further evaluated in future trials and the loss associated with declaring an ineffective schedule acceptable is less than the loss associated with declaring an unsafe schedule acceptable. We could decrease the probability of selecting an unacceptable schedule in this scenario by increasing γ_E.

Table 2.2 presents operating characteristics for our alternative scenarios. The results are broken down into two strata. In the first stratum, Schedule 1 is unacceptable ($\pi_T = 0.25$, $\pi_E = 0.5$), while Schedule 2 is acceptable with $\pi_T = 0.1$ and various combinations of $\pi_E = \{0.65, 0.8, 0.95\}$ and $\mu_2 = \{0, log(2)\}$. In the second stratum, both schedules have a toxicity rate of 0.1 and immune response rate of 0.95 (i.e., nearly perfect toxicity and immune response rates) but we vary μ_2 while leaving $\mu_1 = 0$ to determine the ability of our model to distinguish between schedules that differ only in the magnitude of immune response. In the first stratum of Table 2.2 when only Schedule 2 is acceptable, we correctly identify Schedule 2 as better in 51-67% of all simulated trials after Stage 1, with only a 2.6-3.5% increase after Stage 2. We are unable to reach a conclusive result after stage 1 in 4.3-6.1% of trials and expand to Stage 2 in 3.6-5.3% of all simulated trials. This results in a total expected sample size less than 22 in all cases, suggesting that our design has adequate power for correctly identifying the better schedule with the small sample sizes typically observed in Phase I cancer trials.

We observe a larger benefit from our two-stage approach in the second stratum. In the second stratum, both schedules have the same toxicity and immune response rate $(0.1, 0.95)$ but we vary the magnitude of immune response for Schedule 2 from $\mu_2 = 0$ to $\mu_2 = log(3)$. The probability of correctly identifying the better schedule increases from 47.3%, with $\mu_2 = log(1.5)$, to 69.9% of the time, when $\mu_2 = log(3)$. More importantly, the second stratum illustrates the advantage of our two-stage approach. For example, when $\mu_2 = log(1.5)$, we decrease the chance of equivocal results from 38.1% after stage 1 to 22.5% after Stage 2 and increase the chance of correctly identifying the better schedule from 32.9% after Stage 1 to 47.3% after Stage 2, while only increasing the total sample size by 9.8 patients, on average. Similarly, when $\mu_2 = log(2)$, we decrease the chance of equivocal results from 26.8% after Stage 1 to 11.0% after Stage 2 and increase the chance of correctly identifying the better schedule from 47.8% after Stage 1 to 63.1% after Stage 2, while only increasing the total sample size by 3.2 patients, on average. An increase in power was also observed for $\mu_2 = log(3)$ but the difference was not as dramatic, as the rate of equivocal results after Stage 1 was much lower than the other two cases.

Table 2.1: Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2. Also presented are the probability of stopping the trial for an inconclusive result after stage 1 (Stop Trial), the expected overall sample size, EN = $2 \times (n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2)$, and the average Stage 2 sample size for each schedule conditional on expanding to stage 2, $\bar{n}_2$. The correct outcome is in bold.

(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_1	μ_2	Stage	Outcomes After Stage				Stop Trial	EN	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.			
(0.25,0.5)	(0.25,0.5)	0	0	1	**0.791**	0.089	0.111	0.009			
				2	**0.792**	0.090	0.114	0	0.004	20.2	18
(0.25,0.9)	(0.25,0.9)	0	0	1	**0.607**	0.171	0.160	0.062			
				2	**0.619**	0.187	0.178	0.014	0.002	22.0	17
(0.05,0.5)	(0.05,0.5)	0	0	1	**0.384**	0.244	0.246	0.126			
				2	**0.389**	0.266	0.274	0.042	0.029	23.0	15

Table 2.2: Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2. Also presented are the probability of stopping the trial for an inconclusive result after stage 1 (Stop Trial), the expected overall sample size, EN $= 2 \times (n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2)$, and the average Stage 2 sample size for each schedule conditional on expanding to stage 2, $\bar{n}_2$. The correct outcome is in bold.

(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_1	μ_2	Stage	Outcomes After Stage				Stop Trial	EN	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.			
(0.25,0.5)	(0.1,0.65)	0	0	1	0.386	0.059	**0.512**	0.043			
				2	0.387	0.059	**0.538**	0.009	0.007	21.4	18
(0.25,0.5)	(0.1,0.8)	0	0	1	0.282	0.034	**0.623**	0.061			
				2	0.282	0.037	**0.656**	0.016	0.009	21.8	17
(0.25,0.5)	(0.1,0.95)	0	0	1	0.224	0.045	**0.671**	0.06			
				2	0.225	0.046	**0.706**	0.016	0.007	21.4	13
(0.25,0.5)	(0.1,0.8)	0	log(2)	1	0.254	0.02	**0.694**	0.032			
				2	0.255	0.021	**0.714**	0.003	0.007	20.6	11
(0.1,0.95)	(0.1,0.95)	0	0	1	0.077	0.21	0.24	**0.473**			
				2	0.080	0.245	0.286	0.275	**0.114**	32.8	18
(0.1,0.95)	(0.1,0.95)	0	log(1.5)	1	0.085	0.205	**0.329**	0.381			
				2	0.085	0.217	**0.473**	0.143	0.082	29.8	16
(0.1,0.95)	(0.1,0.95)	0	log(2)	1	0.057	0.197	**0.478**	0.268			
				2	0.057	0.202	**0.631**	0.052	0.058	23.2	15
(0.1,0.95)	(0.1,0.95)	0	log(3)	1	0.069	0.197	**0.631**	0.103			
				2	0.070	0.203	**0.699**	0.014	0.014	21.8	11

We completed additional simulations to evaluate the impact of varying the design parameters specified above on the operating characteristics of our trial. First, we performed additional simulation studies for an increased Stage 1 sample size but keeping the same maximum sample size, $n = 50$ per schedule. We considered $n_1 = 15$ with maximum $n_2 = 35$ and $n_1 = 20$ with maximum $n_2 = 30$ per schedule. Increasing n_1 increased the probability of correctly identifying the better schedule, but came at the cost of substantially increasing the average total sample size. Specifically, we observed a 10-15% increase in the probability of correctly identifying the better schedule, but a 20-30% increase in the total sample size when $n_1 = 15$, and a 25-50% increase in the total sample size when $n_1 = 20$. Furthermore, we observe that increasing n_1 to 20 negates the benefits of the two-stage design because trials that are inconclusive after Stage 1 will likely require a much larger sample size to achieve a conclusive result. From these results, we see that $n_1 = 15$ provides a reasonable trade-off between the power and sample size and would be a reasonable choice; but the trade-off between power and sample size for $n_1 = 20$ was less desirable. We ran smaller simulation studies to see how sensitive our operating characteristics were to changes in γ_T, γ_E, γ_θ and β, changing one parameter at a time. The results can be found in the Supplementary Materials at the end of the chapter. Most notably, increasing β or decreasing γ_θ resulted in a decreased probability of correctly selecting the better schedule at study completion and an increase in the number of trials that stop after Stage 1 with an inconclusive result.

For a prior sensitivity analysis (also presented in Section 2.6, Supplementary Materials), we varied the upper limit for our $Uniform(0, b)$ prior on σ from $b = 1, 2, 5, 10$ and we varied the variance for our Normal$(0, \tau^2)$ prior on μ from $\tau^2 = 1, 3, 10, 100$, assuming the same simulation and design parameters discussed in Section 2.3. These changes had very little impact on the operating characteristics of our two-stage design.

2.4 Extending to K Vaccine Therapies

To this point, we have only considered the case of two vaccination schedules for simplicity. However, we can easily extend our design to K schedules as follows. In Stage 1, we randomize n_1 patients to each of the K schedules. After Stage 1, estimate $\pi_{T,k}$ and $\pi_{E,k}$

for each schedule, k, and determine if any of the K schedules meet the minimum clinical performance levels. After identifying the clinically acceptable schedules, we compare the posterior expected magnitudes of immune response to identify the superior schedule. If a conclusive result cannot be determined in the first stage, we take the top two schedules to Stage 2 and calculate the additional sample size required using Bayesian predictive probabilities, as described in Section 2.2.2. We can implement hierarchical modeling to share response information across schedules and potentially improve estimation (Berry et al., 2010). Instead of fixing the prior mean magnitude of response, $\mu_k \sim N(0, 3)$ for $k = 1, \ldots K$, we allow the average mean magnitude response to be estimated using all K responses, i.e.

$$\mu_k \sim N(\beta, \tau^2)$$
$$\beta \sim N(0, 3); \;\; \tau^2 \sim IG(3, 1),$$

where IG is an inverse gamma distribution. Note the variance parameter, τ^2, controls the amount of sharing across schedules for estimating the mean magnitude of response. Our prior specification for τ^2 suggests little to no sharing across schedules, *a priori*.

We now briefly investigate the performance of our design in the extended case through a small simulation study. We consider four scenarios with $K = 4$ schedules per scenario. In the first scenario, only one schedule meets the minimum performance level, Criteria (2.1); the toxicity and response rates for the four vaccination schedules are $\{(0.1, 0.8), (0.25, 0.5), (0.05, 0.5), (0.35, 0.8)\}$, respectively, with $\mu_k = 0$ for all $k = 1 : 4$. In the second and third scenarios, there are two acceptable schedules, Schedules 1 and 2, with toxicity and immune response rates equal to $(0.1, 0.95)$ and $(0.1, 0.65)$, respectively, but we set $\mu_1 = 0$ in Scenario 2 and $\mu_1 = log(1.5)$ in scenario 3. Finally, in Scenario 4, Schedules 1 and 2 have the same toxicity and immune response rates, $(0.1, 0.95)$, but $\mu_1 = log(2)$, compared to $\mu_2 = 0$. The true toxicity rate, immune response rate and magnitude of immune response for Schedules 3 and 4 remain the same throughout the four scenarios considered.

Results from 1000 simulated trials for Scenarios 1-4 are displayed in Figure 2.1. The white fill represents the probability of reaching each conclusion after Stage 1, the dark fill represents the increase in the probability of reaching each conclusion when the trial

expands to Stage 2, and the diagonal striped fill represents the probability of stopping the trial early after Stage 1; i.e., the sample size required to achieve our desired predictive probability exceeds our pre-defined maximum sample size. Similar to previous results, we do not observe a large benefit in the two-stage design when only one schedule is acceptable. Here, we identify schedule 1 as better in 42.4% of all simulated trials with a 5% increase after Stage 2. Scenarios 2 through 4 illustrate the advantage of our two-stage approach. We are most likely to correctly identify Schedule 1 as the best schedule in all cases, but we observe a larger benefit from our two-stage approach in Scenarios 3 and 4, when schedule 1 has a larger magnitude of immune response. In Scenario 3, we see a 12% increase (from 41% to 53%) in correctly identifying Schedule 1 after Stage 2 while in Scenario 4, we see a 11.7% increase (from 46.3% to 58%). Finally, we note that expanding to Stage 2 dramatically decreases the number of inconclusive results in all cases, which illustrates the advantage of the two-stage approach.

2.5 Discussion

We have presented a Phase I-II trial design for therapeutic cancer vaccines. A unique aspect of this design is that we are comparing vaccination schedules, rather than dose levels, and there is neither a complete nor partial ordering of the vaccination schedules. In this case, we consider a randomized design, where subjects are randomized to one of two or more vaccination schedules, rather than a dose-escalation design, which is the standard approach for Phase I clinical trials in oncology. Dose-escalation designs are used in standard Phase I clinical trials in oncology for ethical reasons (i.e., it is not ethical to treat subjects at higher dose levels when lower dose levels have not been proven safe). In our setting, there are no ethical concerns about randomizing subjects to a vaccination schedule because there is no *a priori* ordering of the schedules. Randomization is the gold standard for comparing two treatment assignments, which allows our design to make an unbiased comparison between the two schedules' responses.

Our design uses a two-stage approach to evaluate and compare vaccination schedules. In the first stage, we evaluate the rate of dose limiting toxicities and the immune response rate and use these results to identify the set of acceptable vaccination schedules.

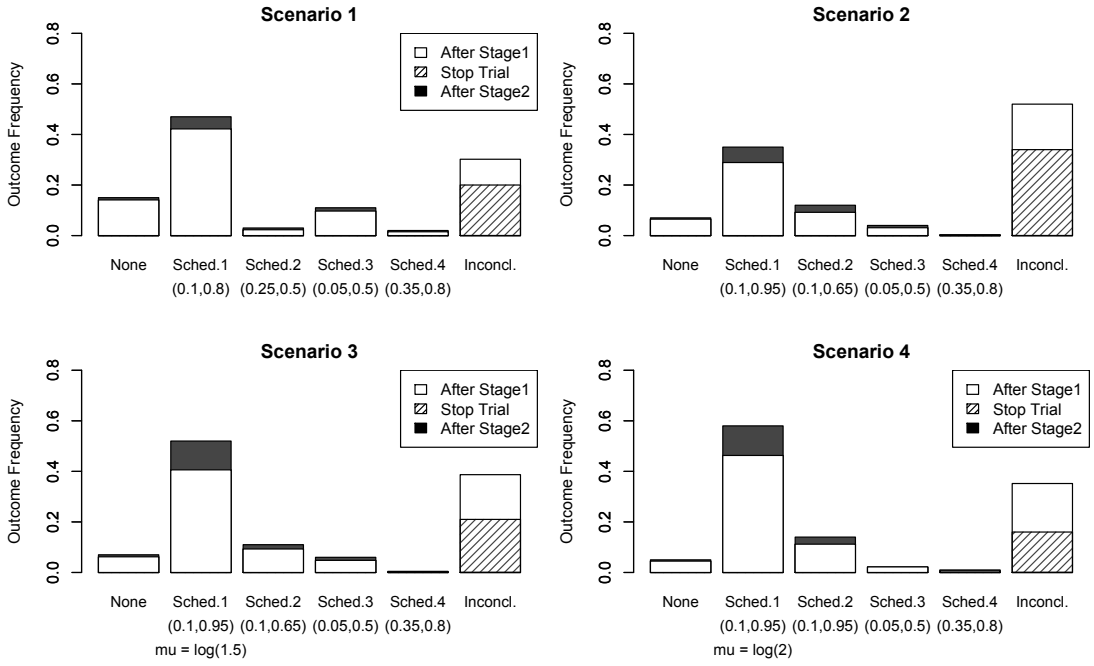

Figure 2.1: Outcome frequency for determining the best schedule when comparing four vaccination schedules. The white fill represents the outcome probabilities after the first stage; the grey fill represents the improved outcome probabilities after the second stage. In the "Inconcl." bar, the diagonal fill represents the percent of trials that stop after Stage 1 because the best schedule could not be determined using predictive probability with the maximum sample size enrolled. True toxicity and immune response rates (and magnitude) are displayed directly below each vaccination schedule.

Ideally, we would be able to make a definitive statement as to the superiority of one of the schedules after the first stage but, if this is not possible, we expand to a second stage to complete a more precise comparison of the schedules. We use Bayesian predictive probabilities to determine the sample size required to achieve a conclusive result and the trial terminates for futility if we are unlikely to reach a definitive conclusion with a pre-determined maximum sample size. Our simulation results show that our two-stage approach dramatically reduces the number of inconclusive trials with only a small increase in the expected sample size. Additional simulation results indicate similar performance when we generalize our design from two to four vaccination schedules under evaluation.

We determine the superior vaccination schedule among acceptable schedules using the magnitude of immune response at the suggestion of our clinical collaborators. We modeled immune response as a zero-inflated log-normal distribution and expected immune response as a function of both the immune response rate and the expected magnitude of immune response for subjects conditional on a non-zero response. Other approaches to identifying the best vaccination schedule could also be considered. For example, the immune response rate could be used to identify the best schedule unless multiple acceptable schedules have 100% immune response, in which case magnitude of response could be used to break the tie. Alternately, a formal decision theoretic approach could be used to consider the trade-off between the rate of toxicity, immune response rate and magnitude of response. We chose to proceed with our approach due to clinical relevance but other approaches should also be investigated.

Typically, the outcomes of a Phase I-II clinical trial are dose-limiting toxicities and tumor response, or some other measure of clinical outcome. In our case, we use the same toxicity outcome but our efficacy outcome is immune response, rather than tumor response. Immune response is an inexact surrogate for clinical response and a subsequent Phase II trial would be needed to evaluate the clinical benefit of the best vaccination schedule identified from this trial. Nevertheless, we feel that "Phase I-II" is an accurate description of our proposed design because we are considering both a toxicity and efficacy endpoint. Future work to incorporate both immune response and tumor

response could follow the approach of Zhong et al. (2012), who developed a trivariate continual reassessment method for incorporating surrogate efficacy into dose-finding.

2.6 Supplementary Materials

Table 2.3: Simulation results when $n_1 = 15$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2. Also presented are the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected overall sample size, EN $= 2 \times (n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2)$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold.

(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_1	μ_2	Stage	Outcomes After Stage				Stop Trial	EN	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.			
(0.25,0.5)	(0.25,0.5)	0	0	1	**0.772**	0.112	0.103	0.013			
				2	**0.777**	0.114	0.108	0.001	0	30.4	14
(0.25,0.9)	(0.25,0.9)			1	**0.586**	0.175	0.183	0.056			
				2	**0.595**	0.182	0.192	0.016	0.015	31.8	22
(0.05,0.5)	(0.05,0.5)			1	**0.274**	0.28	0.244	0.202			
				2	**0.284**	0.326	0.279	0.065	0.046	37	22
(0.25,0.5)	(0.1,0.65)	0	0	1	0.24	0.031	**0.654**	0.075			
				2	0.241	0.033	**0.687**	0.026	0.013	32.4	20
(0.25,0.5)	(0.1,0.8)			1	0.165	0.028	**0.727**	0.08			
				2	0.165	0.029	**0.764**	0.027	0.015	32.6	20
(0.25,0.5)	(0.1,0.95)			1	0.154	0.024	**0.757**	0.065			
				2	0.154	0.024	**0.801**	0.01	0.011	32.2	20
(0.25,0.5)	(0.1,0.8)		log(2)	1	0.163	0.024	**0.788**	0.025			
				2	0.165	0.024	**0.808**	0.001	0.002	30.6	13
(0.1,0.95)	(0.1,0.95)	0	0	1	0.034	0.191	0.21	**0.565**			
				2	0.034	0.225	0.245	0.247	**0.249**	44.4	23
(0.1,0.95)	(0.1,0.95)		log(1.5)	1	0.032	0.15	**0.412**	0.406			
				2	0.032	0.156	**0.563**	0.107	0.142	41.2	21
(0.1,0.95)	(0.1,0.95)		log(2)	1	0.034	0.133	**0.628**	0.205			
				2	0.035	0.134	**0.75**	0.025	0.056	35.6	19
(0.1,0.95)	(0.1,0.95)		log(3)	1	0.031	0.147	**0.763**	0.059			
				2	0.031	0.147	**0.809**	0.003	0.01	31.6	16

Table 2.4: Simulation results when $n_1 = 20$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2. Also presented are the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected overall sample size, EN $= 2 \times (n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2)$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold.

(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_1	μ_2	Stage	Outcomes After Stage				Stop Trial	EN	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.			
(0.25,0.5)	(0.25,0.5)	0	0	1	**0.809**	0.1	0.082	0.009			
				2	**0.81**	0.102	0.083	0.001	0.004	40.2	16
(0.25,0.9)	(0.25,0.9)			1	**0.61**	0.17	0.178	0.042			
				2	**0.61**	0.176	0.179	0.006	0.029	40.4	12
(0.05,0.5)	(0.05,0.5)			1	**0.297**	0.245	0.297	0.161			
				2	**0.301**	0.258	0.307	0.018	0.116	41.2	14
(0.25,0.5)	(0.1,0.65)	0	0	1	0.211	0.033	**0.693**	0.063			
				2	0.211	0.033	**0.71**	0.012	0.034	40.8	14
(0.25,0.5)	(0.1,0.8)			1	0.11	0.018	**0.806**	0.066			
				2	0.11	0.018	**0.821**	0.009	0.042	40.6	11
(0.25,0.5)	(0.1,0.95)			1	0.115	0.01	**0.82**	0.055			
				2	0.115	0.01	**0.836**	0.008	0.031	40.6	13
(0.25,0.5)	(0.1,0.8)		log(2)	1	0.124	0.01	**0.854**	0.012			
				2	0.124	0.01	**0.861**	0.001	0.004	40.2	10
(0.1,0.95)	(0.1,0.95)	0	0	1	0.03	0.166	0.208	**0.596**			
				2	0.03	0.184	0.221	0.095	**0.47**	43.6	14
(0.1,0.95)	(0.1,0.95)		log(1.5)	1	0.013	0.115	**0.461**	0.411			
				2	0.013	0.117	**0.538**	0.074	0.258	44	13
(0.1,0.95)	(0.1,0.95)		log(2)	1	0.013	0.147	**0.687**	0.153			
				2	0.013	0.147	**0.743**	0.016	0.081	42	14
(0.1,0.95)	(0.1,0.95)		log(3)	1	0.016	0.132	**0.827**	0.025			
				2	0.016	0.133	**0.841**	0.001	0.009	40.4	11

Table 2.5: Simulation results when $n_1 = 10$, $n = 50$. Changing: γ_T, γ_E, γ_θ and β (with the no. of stimulated stage 2's used to estimate n_2) is presented in the first column. Presented are: the probabilities of declaring a schedule better after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size/schedule, En= $n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2$, and the average Stage 2 sample size per schedule conditional on expanding to stage 2, $\bar{n}_2$. The correct outcome is in bold.

Parameter	(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_2	Outcomes After Stage				Stop Trial	En	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.		
$\gamma_T = 0.75$	(0.25,0.5)	(0.25,0.5)	0	**0.81**	0.093	0.09	0.007			
				0.813	0.093	0.093	0	0.001	10.1	12
	(0.25,0.9)	(0.25,0.9)		**0.58**	0.166	0.207	0.047			
				0.588	0.181	0.22	0.007	0.004	10.7	17
	(0.05,0.5)	(0.05,0.5)		**0.353**	0.258	0.24	0.149			
				0.361	0.288	0.264	0.061	0.026	11.8	15
$\gamma_E = 0.4$	(0.25,0.5)	(0.25,0.5)	0	**0.727**	0.122	0.13	0.021			
				0.734	0.13	0.133	0.002	0.001	10.3	16
	(0.25,0.9)	(0.25,0.9)		**0.599**	0.179	0.17	0.052			
				0.604	0.198	0.181	0.009	0.008	10.8	18
	(0.05,0.5)	(0.05,0.5)		**0.166**	0.272	0.269	0.293			
				0.176	0.322	0.331	0.12	0.051	13.5	14
$\gamma_\theta = 0.05$	(0.1,0.65)	(0.1,0.95)	0	0.096	0.167	**0.343**	0.394			
				0.096	0.175	**0.406**	0.146	0.177	13.4	16
	(0.1,0.65)	(0.1,0.95)	log(1.5)	0.115	0.131	**0.441**	0.313			
				0.115	0.137	**0.578**	0.073	0.097	12.8	13
$\beta = 60\%(500)$	(0.1,0.65)	(0.1,0.95)	0	0.113	0.165	**0.398**	0.324			
				0.113	0.173	**0.489**	0.12	0.105	13.7	17
	(0.1,0.65)	(0.1,0.95)	log(1.5)	0.121	0.144	**0.503**	0.232			
				0.122	0.152	**0.626**	0.05	0.05	12.6	14
$\beta = 70\%(100)$	(0.1,0.65)	(0.1,0.95)	0	0.12	0.166	**0.376**	0.338			
				0.12	0.17	**0.426**	0.051	0.233	12	19
	(0.1,0.65)	(0.1,0.95)	log(1.5)	0.115	0.162	**0.494**	0.229			
				0.116	0.166	**0.566**	0.026	0.126	11.7	16

Table 2.6: **Alt. Prior:** $\sigma \sim Uniform(0,1)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold.

(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_1	μ_2	Stage	Outcomes After Stage				Stop Trial	En	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.			
(0.25,0.5)	(0.25,0.5)	0	0	1	**0.805**	0.082	0.105	0.008			
				2	**0.808**	0.083	0.108	0.001	0	10.1	10
(0.25,0.9)	(0.25,0.9)			1	**0.573**	0.189	0.2	0.038			
				2	**0.579**	0.2	0.208	0.006	0.007	10.6	18
(0.05,0.5)	(0.05,0.5)			1	**0.388**	0.236	0.248	0.128			
				2	**0.397**	0.26	0.27	0.05	0.023	11.5	15
(0.25,0.5)	(0.1,0.65)	0	0	1	0.368	0.06	**0.521**	0.051			
				2	0.369	0.065	**0.544**	0.019	0.003	10.5	11
(0.25,0.5)	(0.1,0.8)			1	0.297	0.018	**0.631**	0.054			
				2	0.298	0.02	**0.662**	0.016	0.004	10.7	15
(0.25,0.5)	(0.1,0.95)			1	0.222	0.036	**0.684**	0.058			
				2	0.222	0.038	**0.724**	0.009	0.007	10.7	14
(0.1,0.95)	(0.1,0.95)	0	0	1	0.055	0.227	0.254	**0.464**			
				2	0.055	0.271	0.297	0.278	**0.099**	16.1	17
(0.1,0.95)	(0.1,0.95)		log(1.5)	1	0.087	0.192	**0.391**	0.33			
				2	0.088	0.205	**0.503**	0.125	0.079	14.1	16
(0.1,0.95)	(0.1,0.95)		log(2)	1	0.064	0.195	**0.546**	0.195			
				2	0.064	0.199	**0.671**	0.035	0.031	12.4	14
(0.1,0.95)	(0.1,0.95)		log(3)	1	0.066	0.209	**0.663**	0.062			
				2	0.066	0.21	**0.71**	0.007	0.007	10.6	11

Table 2.7: **Alt. Prior:** $\sigma \sim Uniform(0,5)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold.

(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_1	μ_2	Stage	Outcomes After Stage				Stop Trial	En	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.			
(0.25,0.5)	(0.25,0.5)	0	0	1	**0.806**	0.078	0.102	0.014			
				2	**0.809**	0.08	0.108	0.002	0.001	10.1	8
(0.25,0.9)	(0.25,0.9)			1	**0.575**	0.206	0.176	0.043			
				2	**0.584**	0.215	0.183	0.006	0.012	10.6	19
(0.05,0.5)	(0.05,0.5)			1	**0.354**	0.268	0.247	0.131			
				2	**0.364**	0.288	0.278	0.039	0.031	11.4	14
(0.25,0.5)	(0.1,0.65)	0	0	1	0.375	0.048	**0.519**	0.058			
				2	0.377	0.051	**0.544**	0.02	0.008	10.8	17
(0.25,0.5)	(0.1,0.8)			1	0.236	0.047	**0.648**	0.069			
				2	0.237	0.048	**0.687**	0.019	0.009	10.9	15
(0.25,0.5)	(0.1,0.95)			1	0.247	0.025	**0.673**	0.055			
				2	0.248	0.025	**0.701**	0.02	0.006	10.7	13
(0.1,0.95)	(0.1,0.95)	0	0	1	0.07	0.238	0.238	**0.454**			
				2	0.074	0.276	0.279	0.259	**0.112**	16.4	19
(0.1,0.95)	(0.1,0.95)		log(1.5)	1	0.091	0.213	**0.308**	0.388			
				2	0.094	0.226	**0.448**	0.143	0.089	14.7	16
(0.1,0.95)	(0.1,0.95)		log(2)	1	0.081	0.207	**0.46**	0.252			
				2	0.081	0.215	**0.595**	0.064	0.045	12.8	14
(0.1,0.95)	(0.1,0.95)		log(3)	1	0.068	0.209	**0.622**	0.101			
				2	0.068	0.214	**0.696**	0.01	0.012	10.8	9

Table 2.8: **Alt. Prior:** $\sigma \sim Uniform(0, 10)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold.

(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_1	μ_2	Stage	Outcomes After Stage				Stop Trial	En	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.			
(0.25,0.5)	(0.25,0.5)	0	0	1	**0.809**	0.084	0.097	0.01			
				2	**0.81**	0.085	0.101	0.002	0.002	10.1	16
(0.25,0.9)	(0.25,0.9)			1	**0.54**	0.21	0.204	0.046			
				2	**0.546**	0.218	0.213	0.012	0.011	10.6	17
(0.05,0.5)	(0.05,0.5)			1	**0.361**	0.249	0.248	0.142			
				2	**0.371**	0.275	0.265	0.058	0.031	11.6	14
(0.25,0.5)	(0.1,0.65)	0	0	1	0.369	0.058	**0.527**	0.046			
				2	0.369	0.059	**0.552**	0.011	0.009	10.6	15
(0.25,0.5)	(0.1,0.8)			1	0.244	0.029	**0.659**	0.068			
				2	0.246	0.03	**0.694**	0.021	0.009	10.7	13
(0.25,0.5)	(0.1,0.95)			1	0.229	0.023	**0.676**	0.072			
				2	0.231	0.024	**0.719**	0.018	0.008	10.9	15
(0.1,0.95)	(0.1,0.95)	0	0	1	0.07	0.233	0.237	**0.46**			
				2	0.07	0.273	0.276	0.272	**0.109**	16.5	18
(0.1,0.95)	(0.1,0.95)		log(1.5)	1	0.076	0.235	**0.332**	0.357			
				2	0.076	0.242	**0.446**	0.148	0.088	14.4	16
(0.1,0.95)	(0.1,0.95)		log(2)	1	0.092	0.2	**0.46**	0.248			
				2	0.092	0.213	**0.59**	0.053	0.052	12.6	13
(0.1,0.95)	(0.1,0.95)		log(3)	1	0.082	0.201	**0.616**	0.101			
				2	0.084	0.204	**0.683**	0.018	0.011	10.9	9

Table 2.9: **Alt. Prior:** $\mu \sim N(0,1)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold.

(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_1	μ_2	Stage	Outcomes After Stage				Stop Trial	En	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.			
(0.25,0.5)	(0.25,0.5)	0	0	1	**0.769**	0.109	0.112	0.01			
				2	**0.77**	0.113	0.114	0.001	0.002	10.1	11
(0.25,0.9)	(0.25,0.9)			1	**0.579**	0.201	0.17	0.05			
				2	**0.585**	0.225	0.18	0.006	0.004	10.9	19
(0.05,0.5)	(0.05,0.5)			1	**0.352**	0.239	0.252	0.157			
				2	**0.359**	0.271	0.283	0.056	0.031	11.9	15
(0.25,0.5)	(0.1,0.65)	0	0	1	0.378	0.046	**0.513**	0.063			
				2	0.378	0.049	**0.547**	0.014	0.012	10.8	16
(0.25,0.5)	(0.1,0.8)			1	0.258	0.026	**0.652**	0.064			
				2	0.26	0.027	**0.692**	0.012	0.009	10.8	14
(0.25,0.5)	(0.1,0.95)			1	0.231	0.03	**0.674**	0.065			
				2	0.231	0.03	**0.708**	0.017	0.014	10.6	12
(0.1,0.95)	(0.1,0.95)	0	0	1	0.081	0.241	0.248	**0.43**			
				2	0.081	0.275	0.287	0.218	**0.139**	15.8	20
(0.1,0.95)	(0.1,0.95)		log(1.5)	1	0.064	0.216	**0.324**	0.396			
				2	0.064	0.229	**0.485**	0.131	0.091	15	16
(0.1,0.95)	(0.1,0.95)		log(2)	1	0.072	0.189	**0.481**	0.258			
				2	0.073	0.202	**0.628**	0.052	0.045	13.4	16
(0.1,0.95)	(0.1,0.95)		log(3)	1	0.06	0.21	**0.629**	0.101			
				2	0.06	0.216	**0.695**	0.009	0.02	10.9	11

Table 2.10: **Alt. Prior:** $\mu \sim N(0, 10)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold.

(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_1	μ_2	Stage	Outcomes After Stage				Stop Trial	En	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.			
(0.25,0.5)	(0.25,0.5)	0	0	1	**0.794**	0.092	0.097	0.017			
				2	**0.799**	0.092	0.104	0.002	0.003	10.2	14
(0.25,0.9)	(0.25,0.9)			1	**0.568**	0.185	0.2	0.047			
				2	**0.576**	0.194	0.214	0.006	0.01	10.6	17
(0.05,0.5)	(0.05,0.5)			1	**0.372**	0.252	0.249	0.127			
				2	**0.384**	0.272	0.27	0.055	0.019	11.5	14
(0.25,0.5)	(0.1,0.65)	0	0	1	0.385	0.048	**0.515**	0.052			
				2	0.387	0.051	**0.545**	0.007	0.01	10.6	14
(0.25,0.5)	(0.1,0.8)			1	0.25	0.04	**0.64**	0.07			
				2	0.251	0.042	**0.681**	0.022	0.004	11	14
(0.25,0.5)	(0.1,0.95)			1	0.241	0.024	**0.682**	0.053			
				2	0.243	0.024	**0.713**	0.015	0.005	10.6	13
(0.1,0.95)	(0.1,0.95)	0	0	1	0.069	0.246	0.259	**0.426**			
				2	0.069	0.28	0.289	0.271	**0.091**	16.1	18
(0.1,0.95)	(0.1,0.95)		log(1.5)	1	0.065	0.21	**0.35**	0.375			
				2	0.068	0.22	**0.486**	0.157	0.069	14.9	16
(0.1,0.95)	(0.1,0.95)		log(2)	1	0.084	0.183	**0.479**	0.254			
				2	0.085	0.191	**0.638**	0.055	0.031	13	14
(0.1,0.95)	(0.1,0.95)		log(3)	1	0.077	0.201	**0.621**	0.101			
				2	0.077	0.204	**0.699**	0.008	0.012	10.9	10

Table 2.11: **Alt. Prior:** $\mu \sim N(0, 100)$. Simulation results when $n_1 = 10$, $n = 50$, $\gamma_T = 0.70$, $\gamma_E = 0.50$, $\gamma_\theta = 0.10$ and $\beta = 0.60$. Presented are: the probabilities of declaring a schedule the better schedule after Stages 1 and 2, the probability of stopping the trial for equivalence after Stage 1 (Stop Trial), the expected sample size per schedule, En $= n_1 + Pr(\text{continuing onto stage 2}) \times \bar{n}_2$, and the average Stage 2 sample size for each schedule conditional on expanding to Stage 2, $\bar{n}_2$. The correct outcome is in bold.

(π_{T_1}, π_{E_1})	(π_{T_2}, π_{E_2})	μ_1	μ_2	Stage	Outcomes After Stage				Stop Trial	En	$\bar{n}_2$
					None	Sched.1	Sched.2	Inconcl.			
(0.25,0.5)	(0.25,0.5)	0	0	1	**0.774**	0.111	0.106	0.009			
				2	**0.779**	0.113	0.107	0	0.001	10.1	17
(0.25,0.9)	(0.25,0.9)			1	**0.56**	0.184	0.213	0.043			
				2	**0.564**	0.193	0.225	0.011	0.007	10.5	14
(0.05,0.5)	(0.05,0.5)			1	**0.353**	0.253	0.258	0.136			
				2	**0.357**	0.273	0.291	0.056	0.023	11.5	13
(0.25,0.5)	(0.1,0.65)	0	0	1	0.366	0.041	**0.531**	0.062			
				2	0.367	0.045	**0.562**	0.021	0.005	10.7	13
(0.25,0.5)	(0.1,0.8)			1	0.274	0.03	**0.639**	0.057			
				2	0.274	0.03	**0.674**	0.015	0.007	10.7	15
(0.25,0.5)	(0.1,0.95)			1	0.219	0.026	**0.684**	0.071			
				2	0.221	0.026	**0.729**	0.015	0.009	11	15
(0.1,0.95)	(0.1,0.95)	0	0	1	0.073	0.22	0.248	**0.459**			
				2	0.073	0.268	0.28	0.273	**0.106**	16.4	18
(0.1,0.95)	(0.1,0.95)		log(1.5)	1	0.066	0.207	**0.343**	0.384			
				2	0.068	0.218	**0.492**	0.153	0.069	15.2	17
(0.1,0.95)	(0.1,0.95)		log(2)	1	0.074	0.198	**0.477**	0.251			
				2	0.074	0.21	**0.611**	0.064	0.041	12.9	14
(0.1,0.95)	(0.1,0.95)		log(3)	1	0.067	0.202	**0.632**	0.099			
				2	0.067	0.205	**0.704**	0.015	0.009	11	11

Chapter 3

Hierarchical Models for Sharing Information Across Populations in Phase I Dose-Escalation Studies

3.1 Introduction

Researchers are often interested in evaluating the performance of a novel treatment in multiple patient populations. In this case, investigators may be required to complete multiple Phase I trials to determine the MTD for each population if the novel treatment is to be used in combination with background standard-of-care that differs by population. For example, results have recently been reported for Phase I clinical trials of Veliparib, a novel PARP inhibitor, in combination with cyclophosphamide for patients with solid tumors and lymphomas (Kummar et al., 2012), whole abdominal radiation for patients with advanced solid malignancies and peritoneal carcinomas (Reiss et al., 2015), whole brain radiation for patients with brain metastases (Mehta et al., 2015) and cisplatin and etoposide in patients with small cell lung cancer (Owonikoko et al., 2015), each resulting in a different estimated MTD.

Completing separate Phase I trials for each patient population is expensive and time-consuming, while collapsing across populations into a single trial would not be scientifically justified. Furthermore, while it is reasonable to assume that the MTD will vary by population, it is also likely that the results of a Phase I trial in one population would provide information about the MTD in the other populations. This motivates a hierarchical modeling (HM) approach that allows each patient population to have a separate MTD but shares information across populations during dose-finding.

3.1.1 Hierarchical Modeling in Phase I Oncology Trials

HM is a widely used statistical approach for sharing information across populations that has been applied to the analysis of clinical trials in a number of settings, including meta-analysis, multi-center trials, multiple comparisons, variable selection, and subgroup analysis (Sutton et al., 2000; Durbec and Sarles, 1978; George and McCulloch, 1993; Stangl, 1995). While HM has been used extensively in Phase III clinical trials, the statistical literature for HM in early phase clinical trials (i.e. Phases I and II) is limited. In Phase I dose-finding trials, HM has been used to pool information across pharmacokinetic data from healthy volunteers (Patterson et al., 1999) and to improve estimation of the probability of a DLT for combinations of doses of two therapeutic agents (Braun and Wang, 2010). HM has been used more extensively in Phase II clinical trials as an approach to properly model treatment response in the presence of disease subpopulations (Berry et al., 2013; Thall et al., 2003) and to identify personalized cancer treatments in genetically-defined subgroups, most notably in the BATTLE (Kim et al., 2011) and I-SPY 2 trials (Barker et al., 2009).

This chapter discusses HM in the context of Phase I oncology trials. HM has been used in Phase II and III clinical trials to facilitate the borrowing of information across patient populations. Applying HM to Phase I oncology trials will allow borrowing of information across populations but provide flexibility by specifying a different MTD for each population. In addition, we also propose novel dose-finding guidelines for Phase I clinical trials using HM. Standard dose-finding approaches that escalate in cohorts of three will not take full the advantage of sharing information across populations, while dose-adaptation after every patient will be too aggressive and jeopardize patient safety.

The dose-finding guidelines proposed in Section 4.3 allow us to fully realize the potential of HM in Phase I clinical trials, while achieving safety profiles similar to standard Phase I trial designs.

The remaining sections of this chapter proceed as follows. First, Section 3.2 discusses hierarchical extensions to three commonly used dose-toxicity models for the CRM. In Section 3.3, we propose dose-finding guidelines and present our dose-finding algorithm for Phase I oncology trials using HM. In Section 3.4, we present simulation results evaluating the operating characteristics of the proposed design. Finally, Section 3.5 concludes with a brief discussion of our findings and the potential for implementing the proposed methods in practice.

3.2 Dose-Toxicity Models

In this section, we discuss hierarchical extensions of three commonly used dose-toxicity models in Phase I clinical trials. In each case, we define a two-level, Bayesian hierarchical model with population-specific effects that allows borrowing across populations. The following notation will be used throughout this section. Let Y_{ikj} be a binary indicator for a DLT in patient $i = 1, \ldots, n_{kj}$, in population $k = 1, \ldots, K$, at dose level $j = 1, \ldots, D$. Denote $\pi_{kj}(i) = Pr(Y_{ikj} = 1 | Dose = j, Population = k)$ as the probability of a DLT for patient i in population k treated at dose level j. Throughout the remainder of this section, we drop the subscript for patient, i, for simplicity.

3.2.1 Power Model

The first model we consider is an extension of the power model presented by O'Quigley and Shen (1996), which is a one-parameter dose-toxicity model commonly used in Phase I oncology trials. We define our hierarchical power model as follows:

$$\pi_{kj} = p_j^{exp(\alpha_k)} \tag{3.1}$$

$$\alpha_k \sim N(A, \sigma^2)$$

$$A \sim N(0, 2^2) \quad \text{and} \quad \log(\sigma) \sim Unif(-1, 1)$$

Here, $(p_1, \ldots, p_D)$ is a monotonically increasing *prior skeleton*, which describes clinicians' prior belief in the probability of a DLT at each dose level. The prior skeleton is constant across populations and specified in advance. Note α_k is the power parameter for the k^{th} population, A represents the shared mean and σ^2 controls the degree of borrowing across populations. From our prior specification, 99% of A's prior mass falls between -6 and 6, which allows the probability of a DLT to range from 0.01 to 0.99 regardless of $0 < p_j < 1$. This is a commonly used prior specification in traditional dose-finding studies (Yin and Yuan, 2009; Liu et al., 2015). The uniform distribution assumed for $log(\sigma)$ is defined on (-1, 1) with a prior mean of 0 (Gelman et al., 2006). Consequently, σ^2 takes values in (0.14, 7.4). When σ^2 is small, our model suggests homogeneity across populations and encourages borrowing. In contrast, large values for σ^2 suggest heterogeneity across populations, resulting in less borrowing. Finally, we note that the power model has been shown to be equivalent to the hyperbolic-tangent model with a different dose transformation (O'Quigley and Shen, 1996; Paoletti and Kramar, 2009).

3.2.2 Logistic Regression Model

An alternative one-parameter model is a logistic regression model with fixed intercept (Goodman et al., 1995):

$$logit(\pi_{kj}) = -3 + \beta_k \times q_j, \tag{3.2}$$

$$\beta_k \sim N(B, \tau^2)$$

$$B \sim N(1, 2^2) \quad \text{and} \quad \log(\tau) \sim Unif(-1, 1)$$

Here, we fix the intercept to be -3 corresponding to a 5% probability of a DLT by chance, i.e., with no dose. Note β_k is the slope parameter for the k^{th} population, B is the shared mean slope and τ^2 controls the degree of borrowing across populations. In this model, $(q_1, \ldots, q_D)$ are scaled dose-levels that can be specified to reflect the investigator's prior expectation for the probability of DLT at each dose. For the simulation results presented in Section 3.4, we specified $(q_1, \ldots, q_D)$ such that the induced values of π_{kj} are equal to the prior skeleton used in Section 3.2.1 assuming a slope equal to the prior mean of 1. Under our prior specification for B, 99% of B's prior mass is between -5 and 7, which allows our probability of a DLT to range from < 0.001 to

1. In addition, we note that while we expect the slope β_k to be positive, we choose a normal prior for the slope, rather than an exponential or gamma prior, because we have found the normal prior to result in better performance and because it results in a more natural hierarchical extension to the standard model. As with the power model, the parameter controlling the amount of sharing, i.e. τ^2's distribution, is defined on (0.14, 7.4).

Originally, we also considered a two-parameter logistic regression model that allowed each population's dose-toxicity model to have a random intercept and slope. This model required that we specify a bivariate normal model for the correlated intercept and slope. We considered the conditionally-conjugate inverse-Wishart for the covariance prior, as well as the more flexible decomposition prior, $\Sigma = SRS$, discussed by Barnard et al. (2000). We found that this model is overly complex for the small sample sizes in Phase I clinical trials and sensitive to prior input values. Furthermore, it did not improve operating characteristics and it has been shown that the one-parameter logistic model has better performance than the two-parameter logistic model in Phase I clinical trials (Paoletti and Kramar, 2009). Therefore, the two-parameter logistic regression model will not be considered further.

3.2.3 Curve-Free Model

The first two models are parametric and potentially vulnerable to model misspecification. An alternate approach is a curve-free method that directly models the probability of DLT at each dose-level while imposing a monotonicity constraint on the relationship between dose and the probability of DLT without specifying a formal parametric form for the dose-toxicity relationship (Gasparini and Eisele, 2000). Gasparini and Eisele (2000) reparameterize the probability of DLT at each dose-level as follows:

$$\{\theta_1 = 1 - \pi_1, \theta_2 = (1 - \pi_2)/(1 - \pi_1), \ldots, \theta_D = (1 - \pi_D)/(1 - \pi_{D-1})\}.$$

The $\{\theta_1, \ldots, \theta_D\}$ are then given independent priors. If we define θ_j to follow a $Beta$ distribution, the induced joint distribution on the probabilities of DLT is a product-of-beta prior (Gasparini and Eisele, 2000). Extending the product-of-beta prior to hierarchical modeling is unclear and awkward. Instead, we specify our hierarchy on the θs after a logit transformation to facilitate borrowing of information across populations.

A hierarchical model for the curve-free method can be specified as follows:

$$logit(\theta_{kj}) = \gamma_{kj}, \tag{3.3}$$

$$\gamma_{kj} \sim N(\Gamma_j, \nu_j^2),$$

$$\Gamma_j \sim N(c_j, 3^2) \quad \text{and} \quad \log(\nu_j) \sim Unif(-1, 1),$$

where γ_{kj} is the unrestricted model parameter for population k and dose j, Γ_j is the shared population mean for dose j and ν_j^2 controls the amount of borrowing across populations for each dose level j, $j = 1, \ldots, D$. As with the logistic regression model discussed in Section 3.2.2, the hyper-parameters (i.e., c_j, $j = 1, \ldots, D$) can be specified such that they induce values of π_{kj} equal to the prior skeleton used in Section 3.2.1. This prior specification implies a 99% prior probability that Γ_j is within $c_j \pm 9$ for $j = 1, \ldots, D$ and, again, the prior distribution for ν_j^2 is defined on (0.14, 7.4).

The prior distributions discussed above represent the final priors used to fit these models but other prior distributions, particularly for the hierarchical variance parameter, were also considered. Specifically, we also considered the conditionally-conjugate inverse-Gamma prior for the variance and a Uniform(0,b) prior on the standard deviation. However, the former is sensitive to prior input values when the estimated standard deviation is small, which can occur early in the trial or for homogeneous populations (Gelman et al., 2006). A Uniform(0,b) prior on the standard deviation places more density on more extreme prior σ^2 values than the Uniform(-a, a) prior on the log standard deviation. As a result, convergence and identifiability are a concern with the small sample sizes found in Phase I clinical trials.

3.3 Dose-Finding Algorithm

In this section, we discuss dose-finding when using hierarchical modeling to share information across populations in Phase I oncology trials. We expect that enrollment will be staggered and randomly distributed across patient populations with several consecutive patients enrolled in one population, while long stretches may occur without enrolling patients in others. As a result, extending standard Phase I dose-finding algorithms to

our case is not trivial. Typically, Phase I dose-escalation studies use cohorts of three patients and the MTD is re-evaluated for each new cohort. There are two natural extensions of this approach but neither is satisfactory. First, we could re-evaluate the MTD after each cohort of three patients, regardless of patient population. This approach would be too aggressive and could result in a patient being treated at a higher dose level before lower dose levels had have been tried in that patient's population. The second option would be to use cohorts of three patients within a patient population and only re-evaluate the MTD within a patient population when a new cohort is ready to enroll. This approach would be too conservative, failing to take full advantage of HM and treat too many patients at sub-therapeutic dose levels. Therefore, we propose three dose-finding guidelines (DFGs) and compare the performance of each through simulation.

Throughout the rest of this section, we select the MTD and terminate the trial for excess toxicity as follows. Let $\overline{\pi}_T$ be a pre-defined target toxicity rate. Dose-level j in population k is considered to have acceptable toxicity if:

$$Pr(\pi_{kj} < \overline{\pi}_T | Data, Dose) > \gamma. \tag{3.4}$$

This is a commonly used criterion in Phase I oncology trials and γ is typically chosen between 0.05 and 0.20. The threshold γ can be thought of as a tuning parameter, which is chosen to achieve the desired operating characteristics for the trial (Berry et al., 2010). Alternately, γ can be considered as the acceptable amount of error, since a dose is only declared unacceptable when the posterior distribution strongly suggests it is overly toxic. We have defined $\overline{\pi}_T$ and γ to be constant across populations but these could be made population-specific, if desired. In the context of the DFGs described below, the MTD for population k, MTD_k, is defined as the dose-level that minimizes the absolute difference between the probability of a DLT and $\overline{\pi}_T$ from among the set of doses with acceptable toxicity. Dose-finding for a population terminates if the lowest dose-level has unacceptable toxicity by Criterion (3.4).

3.3.1 Dose-Finding Guidelines

We now describe three DFGs for Phase I clinical trials with multiple patient populations. In each DFG, initial patients in each population start at the lowest dose-level. The DFGs define when it is acceptable to escalate, at which point dose assignment will depend on the estimated MTD_k for each population $k = 1, \ldots, K$ using the hierarchical models described in Section 3.2.

The most common approach to dose-finding in Phase I oncology trials is to escalate in cohorts of three patients. The first approach we consider is to allow escalation within a population when at least 3 patients have been treated across all populations and at least 1 patient has been treated in the current population. We refer to this approach as the "1(m)" DFG. Here, we describe our DFGs in terms of a general cohort size, m, to allow for additional flexibility as cohort sizes of two and four patients have also been proposed for Phase I oncology trials. Formally, the "1(m)" DFG allows escalation to dose-level $j + 1$ for population $k = 1, \ldots, K$ if:

- **m** patients overall (and at least 1 patient in population k) have been treated at dose level j.

We note that a special case of the "1(m)" DFG occurs when all m patients previously treated at dose-level j are in population k, which would allow escalation and is consistent with the standard Phase I design that escalates in pre-specified cohorts of size m. For example, if $m = 3$, the "1(m)" DFG allows the fourth patient enrolled, say in population k, to potentially escalate to dose-level 2 as long as one of the first three patients treated at dose-level 1 was in population k. This suggests observing m patients within a population is equivalent to observing m patients overall in estimating each population's dose-response curve.

It is possible that the "1(m)" DFG might be too aggressive and result in an unacceptably high number of toxicities. Therefore, we consider two other DFGs with further restrictions to protect patient safety. First, we consider the "m/1($m + 1$)" DFG. The "m/1(m+1)" DFG allows escalation to dose-level $j + 1$ for population $k = 1, \ldots, K$ if:

- **m** patients in population k have been treated at dose level j,

- Or $\mathbf{m+1}$ patients overall (and at least 1 patient in population k) have been treated at dose level j.

For $m = 3$, the "$m/1(m+1)$" DFG allows dose escalation in population k if $m = 3$ patients in population k have been treated at population k's current dose level or if $m + 1 = 4$ patients overall have been treated at population k's current dose level with at least one patient in population k. This suggests observing m patients within a population is equivalent to observing $m + 1$ patients overall in informing our dose-response models.

Finally, we propose a third DFG, which we refer to as the "321" DFG, which is similar to the "$m/1(m+1)$" DFG but puts additional restrictions on escalation to protect patient safety and promote reasonable sharing across populations. The "321" DFG allows escalation to dose-level $j + 1$ for population $k = 1, \ldots, K$ if:

- **3** patients in population k have been treated at dose-level j,
- OR **2** patients in population k and at least 2 patients not in population k have been treated at dose-level j,
- OR **1** patient in population k and at least 1 patient in three other populations have been treated at dose-level j.

The "321" DFG is the most restrictive of the three DFGs. Examples of scenarios where the "$m/1(m+1)$" DFG would allow escalation to dose-level $j + 1$ but the "321" would not include:

- only 1 patient in population k and 3 patients in population k_2 have been treated at dose level j,
- only 1 patient in population k, 2 patients in population k_2 and 1 patient in population k_3 have been treated at dose level j.

The "321" DFG restricts the scenarios where escalation is allowed after only one patient in a population has been treated at the current dose-level to reduce the influence of other populations' estimated dose-response curves, should they be different. While

this restriction slows dose-finding by requiring more patients to enroll, it is incorporated for patient safety.

Figures 3.1 through 3.4 below present a single simulated trial using the three DFGs described above, along with a trial with no restrictions on escalation except that untried dose-levels cannot be skipped when escalating within a population (no DFG). We note that the DFGs defined above indicate when it is *potentially* acceptable to escalate to an untried dose within a population but the ultimate decision to escalate will be based on the current estimate of the MTD_k using all available data. In some sense, the DFGs define a run-in period for each population and dose-level to prevent escalation without sufficient evidence that the current dose-level is safe within each population.

In summary, our dose-finding algorithm for identifying each MTD_k for $k = 1, \ldots, K$ in a Phase I clinical trial using HM proceeds as follows:

1. Treat the first patient within each population at the lowest dose level.
2. When a new patient is enrolled, update the posterior distribution of the probability of toxicity for all dose-levels and populations using all available data.
3. Identify the set of acceptable doses for each population using Criterion (3.4).
4. If all doses are unacceptably toxic and at least 3 patients have been treated in the current population, then the trial terminates for that population. For each population, if all doses are unacceptably toxic but less than 3 patients have been treated in that population, then the next patient is treated at dose-level 1.
5. Otherwise, treat the next patient at the dose-level that minimizes $|E(\pi_{kj}|Data) - \overline{\pi}_T|$ from among the acceptable dose-levels under the restriction that escalation is only allowed if the criterion for escalation is satisfied for the pre-specified DFG and escalating more than one dose-level at a time is not allowed.
6. Repeat steps 2-5 until the maximum overall sample size is reached. Within each population, the acceptable dose that minimizes $|E(\pi_{kj}|Data) - \overline{\pi}_T|$ at study completion is considered the MTD_k for $k = 1, \ldots, K$.

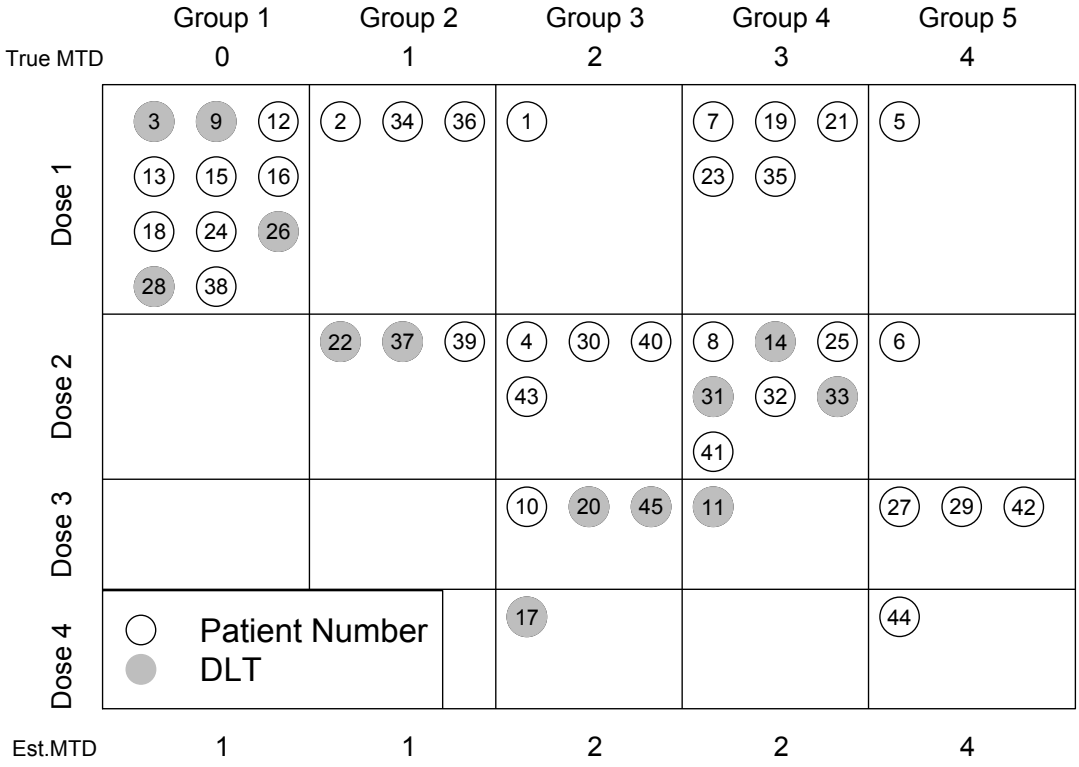

Figure 3.1: **no DFG**: Display of dose-finding for a simulated trial when no DFGs are implemented in our dose-finding algorithm. White circles display the patient number; grey circles depict if the patient experienced a DLT. The true MTD_k is displayed at the top of the figure, directly below each population; the estimated MTD_k is displayed at the bottom of the figure for each population.

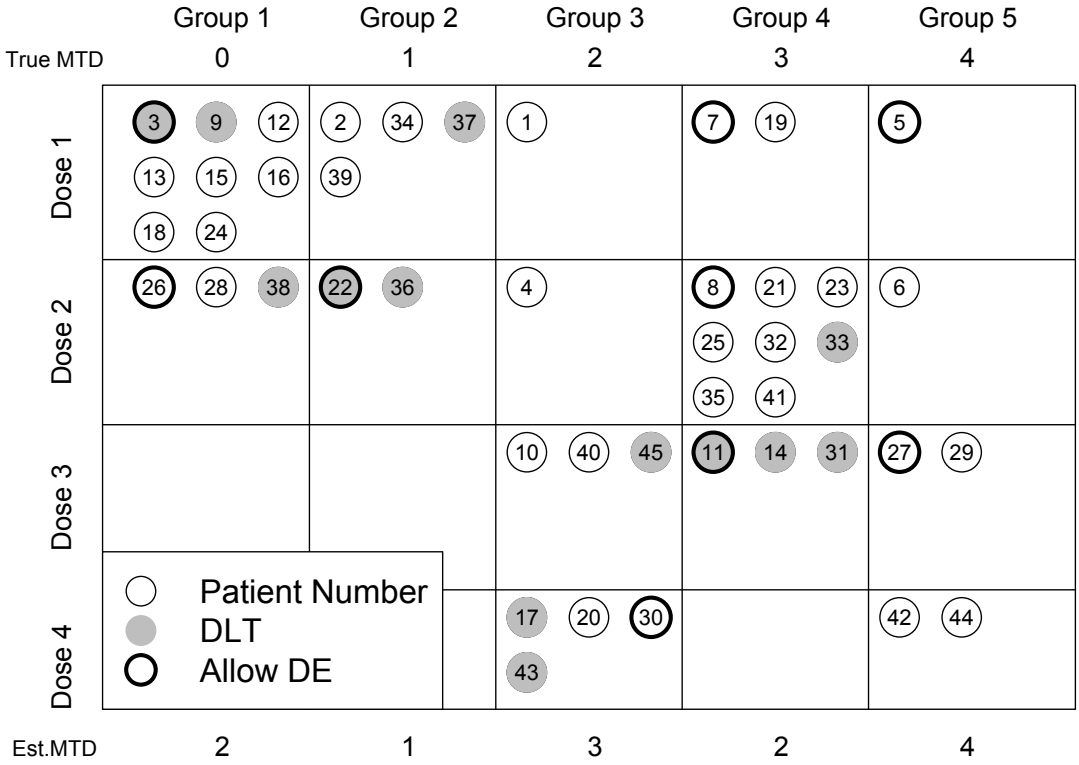

Figure 3.2: **1(m)**: Display of dose-finding for the same simulated trial when the "1(m)" DFGs are implemented in our dose-finding algorithm. White circles display the patient number; grey circles depict if the patient experienced a DLT; bolded circles represent if dose-escalation is allowed for any of the populations after patient enrollment. The true MTD_k is displayed at the top of the figure, directly below each population; the estimated MTD_k is displayed at the bottom of the figure for each population.

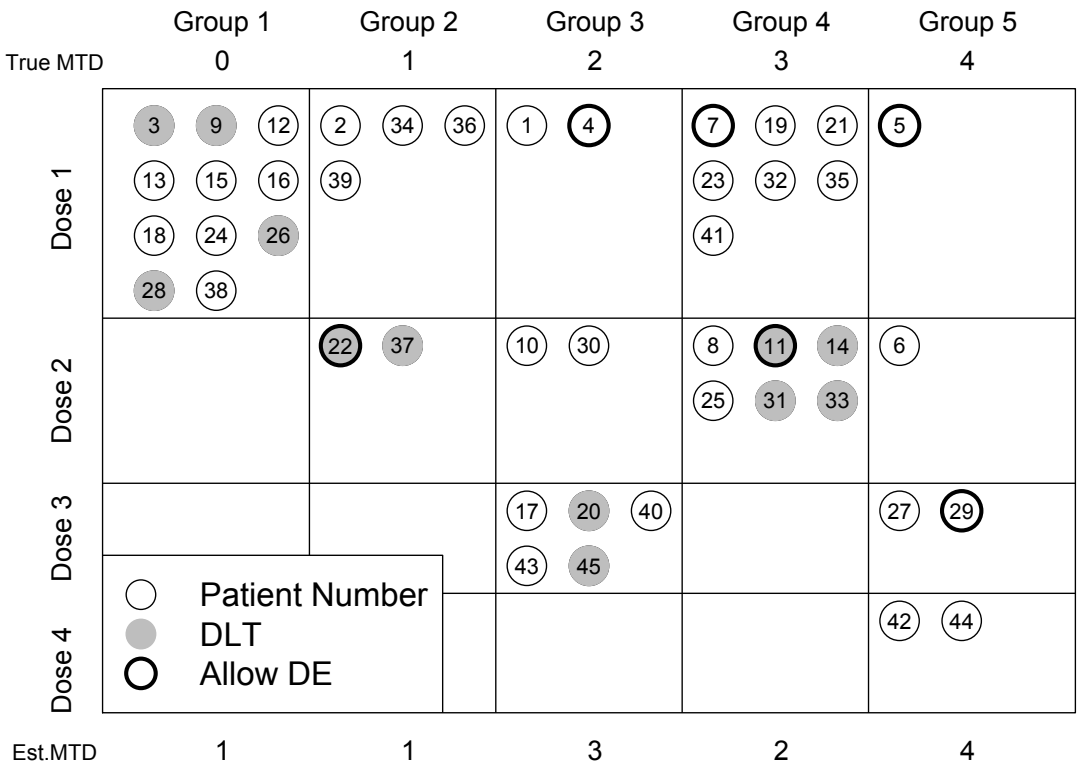

Figure 3.3: **m/1(m+1)**: Display of dose-finding for the same simulated trial when the "m/1(m+1)" DFGs are implemented in our dose-finding algorithm. White circles display the patient number; grey circles depict if the patient experienced a DLT; bolded circles represent if dose-escalation is allowed for any of the populations after patient enrollment. The true MTD_k is displayed at the top of the figure, directly below each population; the estimated MTD_k is displayed at the bottom of the figure for each population.

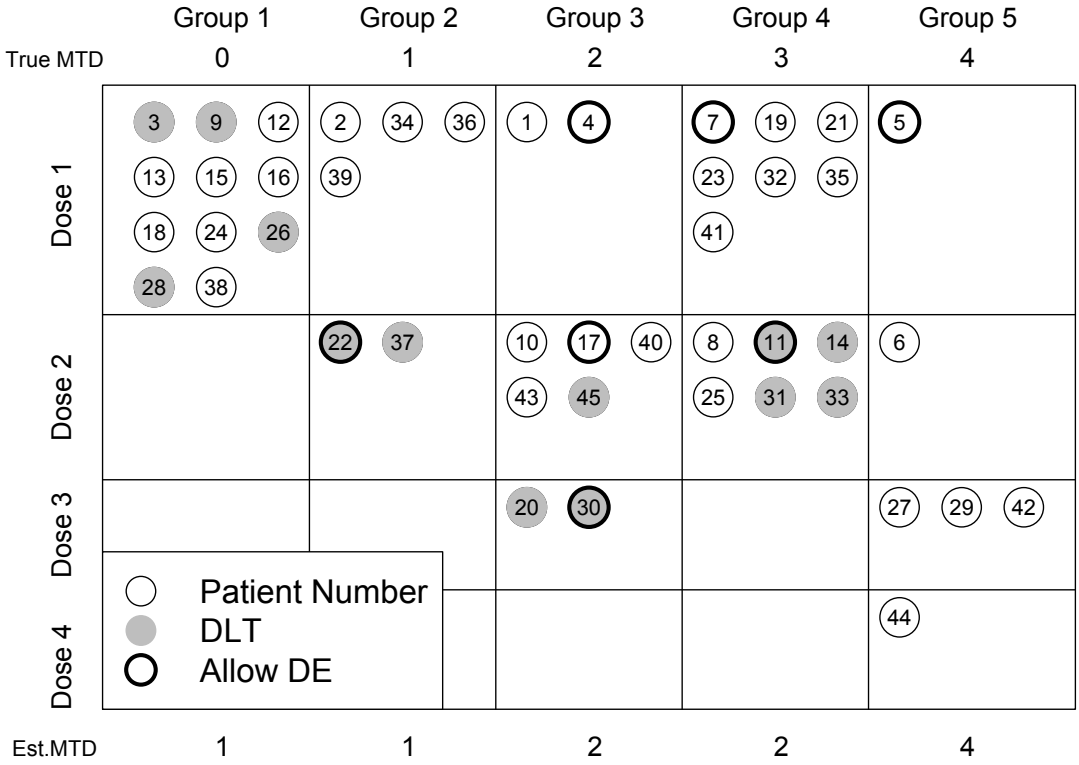

Figure 3.4: **321**: Display of dose-finding for the same simulated trial when the "321" DFGs are implemented in our dose-finding algorithm. White circles display the patient number; grey circles depict if the patient experienced a DLT; bolded circles represent if dose-escalation is allowed for any of the populations after patient enrollment. The true MTD_k is displayed at the top of the figure, directly below each population; the estimated MTD_k is displayed at the bottom of the figure for each population.

We note that, in step 4, only the most recently enrolled and treated population can terminate for excess toxicity (i.e., a population cannot terminate based on an outcome in another population). In addition, we do not force balance across populations by specifying a maximum number of subjects per population. Enforcing balance would likely improve the operating characteristics of the trial but would also dramatically increase the duration of the trial. Our dose-finding algorithm proposes to update the joint posterior for all populations after every patient. We also explored an approach where the model is only updated when the enrolled population is permitted to dose-escalate. The only practical difference between the two approaches is that we allow de-escalation regardless of whether or not a population is allowed to escalate according to the DFG. We feel that this is appropriate because it is always preferable to treat patients at the current estimate of the MTD and only place restrictions on escalation to protect patient safety, which is not a concern when de-escalating. The operating characteristics of this design were very similar to the proposed algorithm.

3.4 Simulation Study

We conducted a simulation study to evaluate the operating characteristics of a Phase I clinical trial using the hierarchical models and DFGs described in Sections 3.2 and 3.3. We evaluate the performance of each method based on (i) the probability of correctly identifying each MTD_k, (ii) the percent of patients experiencing a DLT, (iii) the percent of patients treated at each MTD_k and (iv) the percent of patients treated above each MTD_k for $k = 1, \ldots, K$. Trial parameters were set as follows. We assume a maximum of 45 patients across $K = 5$ populations and assume an equal probability of enrollment to each population (i.e., an average of 9 patients/population). The target toxicity rate was set to $\overline{\pi}_T = 0.3$. The threshold for determining if a dose has an acceptable probability of toxicity, γ, was set equal to 0.2. We consider $D = 4$ dose levels with dose index $\{1, 2, 3, 4\}$. All simulations were completed in R version 3.1.1. Gibbs and slice sampling were completed in JAGS via R using *rjags* (Plummer, 2011). Posterior inference was completed using 10,000 MCMC samples following a period of 5,000 iterations for burn-in. 1000 simulated trials were completed for each scenario.

The skeleton for the power model, $(p_1, \ldots, p_4)$, was set equal to $(0.1, 0.2, 0.35, 0.50)$. The scaled dose-levels for the logistic regression model, $(q_1, \ldots, q_4)$, were set equal to $(logit(0.1) + 3, logit(0.2) + 3, logit(0.35) + 3, logit(0.5) + 3)$ to achieve a prior skeleton for the probability of toxicity equal to the power model skeleton assuming a slope equal to the prior mean of 1 (Sweeting et al., 2013). We similarly set the hyper-parameters for the curve-free method, $(c_1, \ldots, c_4)$, equal to $(1, 1.5, 2, 2.5)$, which induces $(\pi_{k1}, \ldots, \pi_{k4})$ equal to the power model skeleton.

In addition to simulating a Phase I clinical trial using the models and DFGs described in Sections 3.2 and 3.3, we also evaluated the operating characteristics of three other designs, for comparison. First, we considered a Phase I clinical trial using hierarchical modeling that treats each patient at the current estimate of each MTD_k with no restrictions on dose-escalation other than untried dose-levels within a population cannot be skipped when escalating (no DFG). The DFGs described in Section 3.3 were proposed, primarily, to protect patient safety and a comparison to a design with no restrictions on dose-finding will allow us to isolate the impact of the dose-finding guidelines on the various operating characteristics of the trial. In addition, we simulated two types of Phase I designs that did not use hierarchical modeling and fit independent models for each population. For these two designs, we fit the models specified in Section 3.2 without the second level of the hierarchy and specified independent $N(0, 2^2)$ priors for the α_ks, independent $N(1, 2^2)$ priors for the β_ks and independent $N(c_j, 3^2)$ for the γ_{jk}s with c_j as specified in the previous paragraph. For the independent models, we considered a design with a maximum sample size of 45 patients and a design with a maximum sample size of 90 patients, both with dose-adaptation occurring after every patient. This will allow us to evaluate the benefit of using HM, as compared to designs that treat each population as independent data.

3.4.1 Scenarios

We simulated data from the six scenarios presented in Figure 3.5. The true dose-response curves were set by specifying the slope and MTD in a logistic regression model, with the exception of Group 5 in Scenario 6, which is taken to be qualitatively different from the other groups. In Scenario 1 (top, left plot), all five groups have the same dose-toxicity

curve and the optimal design would be to collapse the five groups and complete a single trial assuming a common dose-toxicity curve for all groups. In Scenario 2 (top, right plot), all five groups have the same MTD but different dose-toxicity curves. Scenarios 3 (middle, left plot) and 4 (middle, right plot) represent scenarios where the dose-toxicity curves vary by population with true MTDs ranging from dose 1 to dose 4 in Scenario 3 and dose 1 to dose 3 in Scenario 4. Scenario 4 was suggested to us by a researcher in the field as a particularly difficult case that we would expect to see in practice. In Scenario 5 (bottom, left plot), all doses are unacceptably toxic for group 1, while dose 1 is the MTD for the second and third groups and dose 2 is the MTD for the last two groups. Finally, in Scenario 6 (bottom, right plot) all doses are unacceptably toxic for group 1, while each of the other four groups has a different MTD.

3.4.2 Results

Figure 3.6 presents the probability of correctly identifying each MTD_k and the percent of patients experiencing a DLT for Scenarios 1 through 3 using the power model. Results for Scenarios 4 through 6 can be found in Figure 3.9 in Section 3.6. In Scenario 1, all groups have the same dose-toxicity curve, while in Scenario 2, all groups have the same MTD but different dose-toxicity curves. We see that, in both scenarios, the hierarchical modeling approach results in a higher probability of correctly identifying each MTD_k than the independence model with a maximum sample size of 45 subjects regardless of the DFG. We note that this is a relatively high bar in that the independence model with 45 patients allows adaptation after every subject and does not include the typical restrictions (cohorts of 3, etc.) that are put in place to protect patient safety but may decrease the probability of correctly identifying the MTD. In addition, the hierarchical power model does as well as, or better than in some cases, than the independence design with 90 patients. This indicates that, in Scenarios 1 and 2, hierarchical modeling was as valuable as doubling the maximum sample size from 45 to 90 patients in independent designs. In Scenarios 3 through 6, where the MTD varies across groups, the hierarchical approach improves upon the independence model with a maximum sample size of 45 in most cases, with the exception of Group 1 in Scenarios 5 and 6 and Group 5 in Scenarios 3 and 6. These are scenarios where either all of the doses were unacceptably

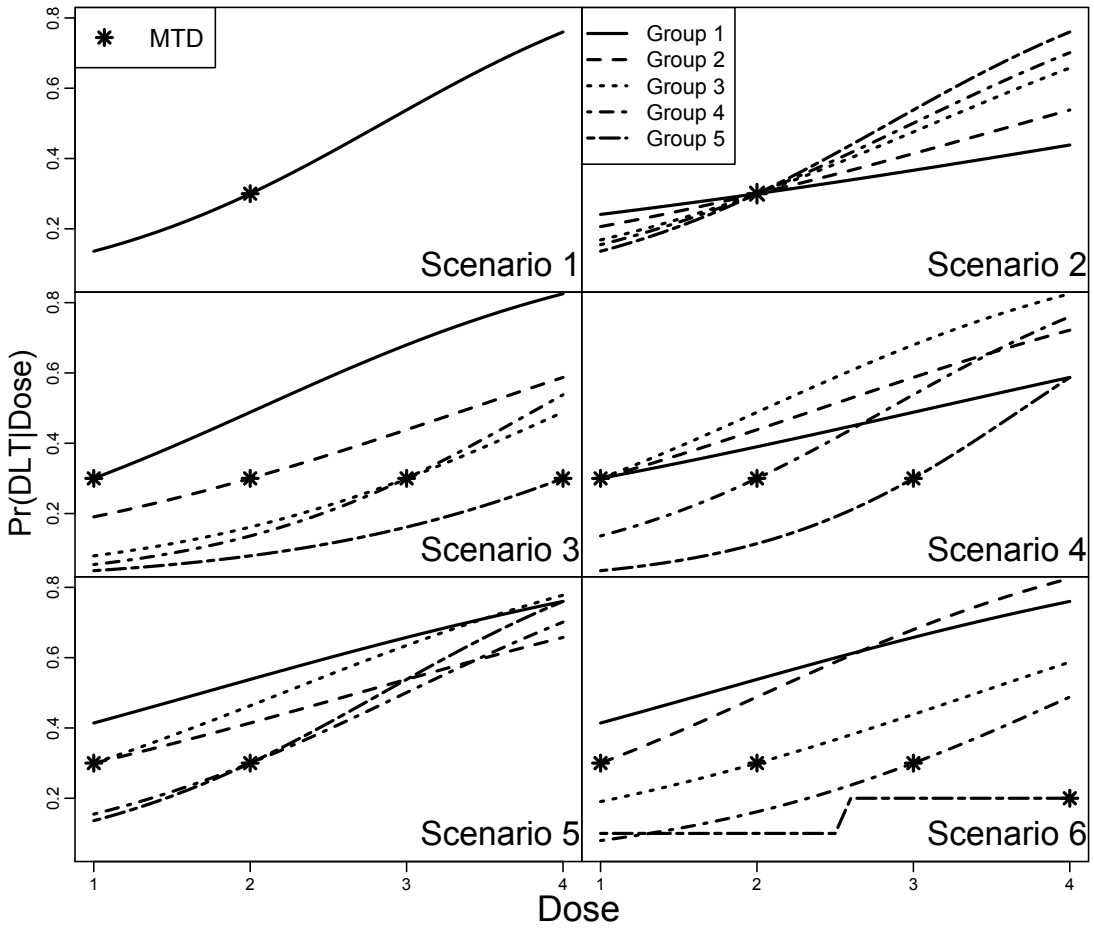

Figure 3.5: Scenario 1 (top, left): all populations' dose-response curve are equivalent. Scenario 2 (top, right): all populations have the same MTD level, but slope increases with population index. Scenario 3 (middle, left): the MTD for each population is dispersed across all four dose-levels for the five populations. Scenario 4 (middle, right): the first three populations' MTD is dose level 1; the last two populations have MTD at dose level 2 and 3, respectively. Scenario 5 (bottom, left): population 1 terminates trial; the next two populations' MTD is dose level 1 the last two populations' MTD is dose level 2. Scenario 6 (bottom, left): similar to Scenario 4, except population 1 terminates trial. The MTD for each population is identified with an asterisk.

toxic or where dose 4 was the MTD_k, suggesting that the hierarchical model was incorrectly shrinking the estimated MTD_k towards intermediate doses due to the results of the other populations. Finally, we note that the three hierarchical DFGs correctly identified each MTD_k at a similar rate as the hierarchical model with no DFG. The goal of implementing the DFGs was to protect patient safety and it is encouraging to see that the three DFGs did not adversely impact the probability of correctly identifying each MTD_k.

Our initial results indicate that HM increases the probability of correctly identifying each MTD_k but a potential concern related to the implementation of this approach is that HM would increase the number of DLTs due to sharing information across populations. Our results suggest that HM does not substantially increase the rate of DLTs and, in fact, results in a decreased probability of DLT relative to the independence designs, in most cases, with the only exceptions occurring when the lowest dose is the MTD for one population but the MTD is higher for other populations. In this case, HM shrinks the estimated MTD_k towards intermediate dose-levels due to the results for the other populations, although even in these cases, the increase in the DLT rate is minor. In addition, we note that our results present the percent of patients experiencing a DLT, rather than the absolute number of DLTs, which implies that the independence model with a maximum sample size of 90 subjects will have a much larger total number of DLTs than the HM designs. Therefore, while the independence design with 90 patients is more likely to accurately identify each MTD_k, in some cases, this comes at the expense of a dramatically higher number of DLTs. Finally, while the differences are subtle, we note that the "321" DFG results in fewer DLTs than the "m/1(m+1)" DFG, which in turn results in fewer DLTs than the "1(m)" DFG, as expected.

Figures 3.10 and 3.11 in Section 3.6 present the average number of patients treated at each MTD_k and the average number of patients treated above each MTD_k. Our results indicate that in addition to increasing the probability of correctly identifying each MTD_k and decreasing the rate of DLTs, HM increases the average number of subjects treated at each MTD_k and decreases the number of patients treated at unsafe doses above each MTD_k, in many cases. Again, exceptions to this rule occur when the lowest

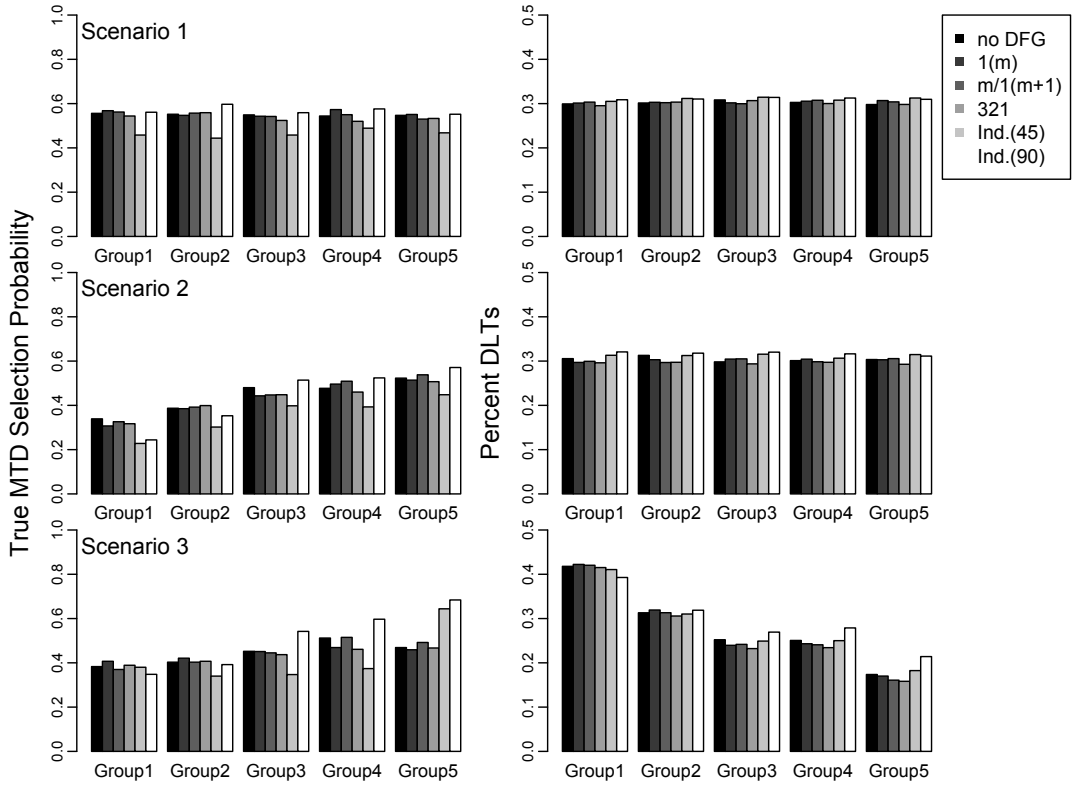

Figure 3.6: **Power Model**: Scenarios 1-3(left plots): Probability of correctly identifying the true MTD; Scenario 1-3 (right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis.

dose is the MTD or when all doses are excessively toxic. In this case, the independence design treats fewer patients above the true MTD_k due to its propensity to stop the trial early and declare all doses overly toxic. We note that results for the independence design with a maximum sample size of 90 subjects were not included in these results because the larger sample size skews our results with respect to the y-axis and because the larger design results in a substantial increase in DLTs compared to the HM designs.

Figure 3.7 and Figures 3.12 through 3.14 in Section 3.6 present results using the hierarchical logistic regression model. Similar to the power model, the logistic regression model performs well when the populations exhibit homogeneity with regards to the MTD (Scenarios 1 and 2), identifying each MTD_k as often as the independence design with a maximum sample size of 90 subjects. Unlike the power model, the hierarchical logistic regression model performed poorly when the MTD varied by population, with particularly poor performance in Scenarios 5 and 6, identifying each MTD_k no more often, and in many cases less often, than the design that assumes independence. Furthermore, we see that, in many cases, the hierarchical logistic regression model treated no more, and in some cases less, patients at the true MTD_k than the independence models, although we note that the logistic regression model was more likely to under-dose.

Finally, results for the hierarchical curve-free model can be found in Figure 3.8 and Figures 3.15 through 3.17 in Section 3.6. This model is non-parametric and thus richly parameterized compared to the other two models. When the groups are homogeneous, we see the HM designs out-perform the two independence designs. On the other hand, the hierarchical curve-free model performed very poorly in some cases. In particular, the hierarchical curve-free model was unlikely to properly identify the MTD_5 for Group 5 in Scenarios 3, 4 and 6. In these cases, Group 5 had a higher MTD_5 than the other groups and the hierarchical component of the model shrank the estimated MTD_5 towards intermediate doses and underestimated the MTD. In addition, this model also performed poorly when all doses were excessively toxic (Group 1 of Scenarios 5 and 6). In a sense, these results are to be expected for this model. In Scenarios 1 and 2, where the populations are homogeneous and the model borrows strengths across groups, the added flexibility of the curve-free method results in a very high probability of correctly

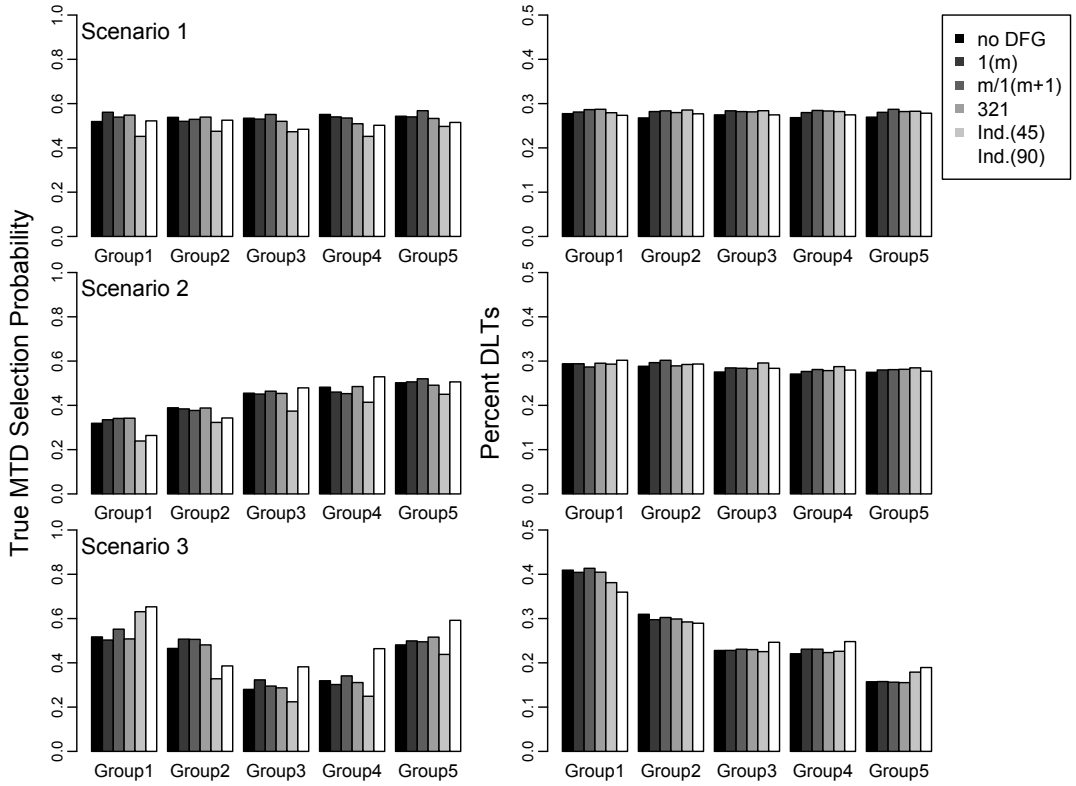

Figure 3.7: **Logistic Regression Model**: Scenarios 1-3(left plots): Probability of correctly identifying the true MTD; Scenario 1-3 (right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis.

identifying each MTD_k. In contrast, when the populations are heterogeneous and there is little borrowing, the model is over-parameterized and is not able to accurately estimate each MTD_k.

In summary, our simulation results suggest that implementing HM in Phase I oncology trials increases the probability of correctly identifying each MTD_k by borrowing information across populations. In addition, HM increases the number of patients treated at each MTD_k, decreases the percent of patients treated at unsafe dose-levels above each MTD_k and decreases the number of toxicities, in most cases. All three models borrowed strength when the MTD was constant across populations, resulting in a more precise estimate of each MTD_k, but the hierarchical power model was more flexible and exhibited better performance when the populations were heterogeneous. In addition, all three DFGs achieve the stated goal of increased patient safety by restricting dose-escalation but the "321" exhibited the best performance with limited impact on the probability of correctly identifying each MTD_k.

3.4.3 Exploring Other K

Our simulation used five populations to evaluate the operating characteristics of a Phase I clinical trial using HM with the models and DFGs specified in Sections 3.2 and 3.3. We now present additional simulation results to evaluate the impact of varying K and determine the minimum number of populations needed to observe a benefit from using HM. Simulation results are presented for $K = 2, 3, 4$ and we only considered the hierarchical power model due to its superior performance in our initial simulation results. The "321" DFG was used for $K = 4$ but is inappropriate for $K < 4$, so results for $K = 2$ and $K = 3$ are presented for the "m/1(m+1)" DFG, instead. Simulations were completed using the prior distributions specified in Section 3.2. However, performance could potentially be improved by reducing the prior domain specified for σ. This is not unreasonable, since we have fewer populations and therefore do not need as large of a domain for σ to produce the appropriate amount of smoothing. Finally, our simulation results assume a maximum sample size of 18, 27, 36 patients for K equal to 2, 3 and 4, respectively, to achieve an average sample size of 9 patients per population, similar to our results with $K = 5$.

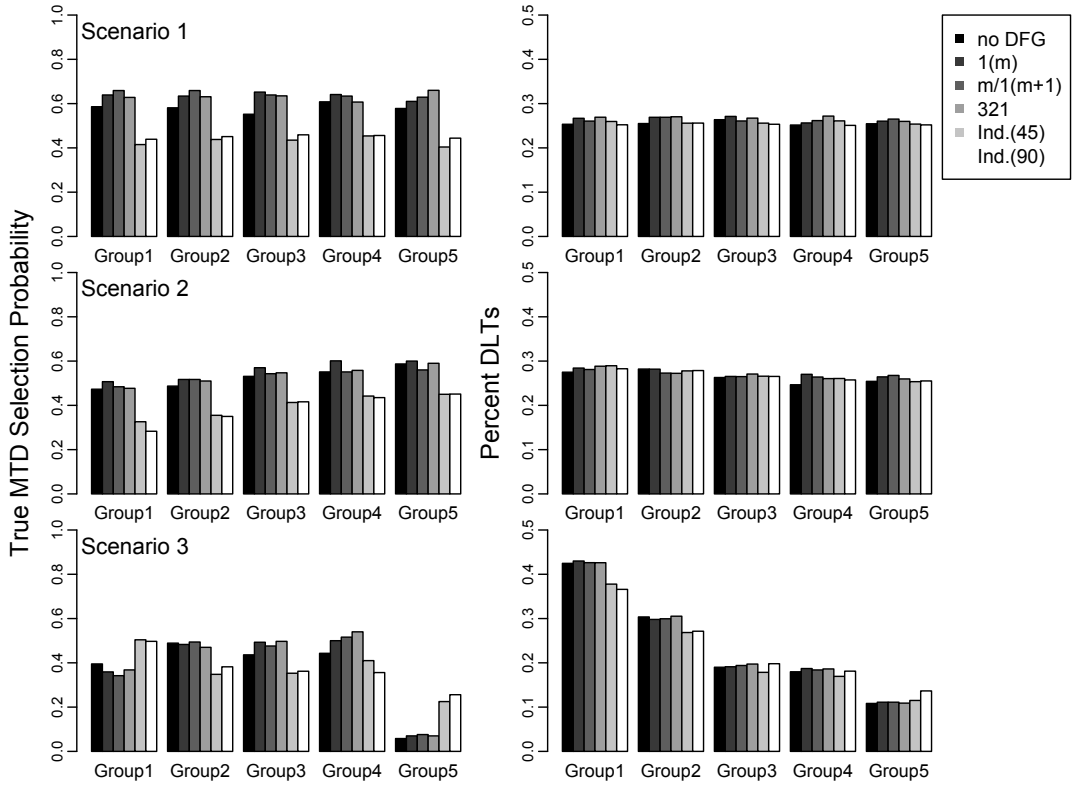

Figure 3.8: **Curve-Free Model**: Scenarios 1-3(left plots): Probability of correctly identifying the true MTD; Scenario 1-3 (right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis.

Simulation results for 1000 simulated trials per scenario are presented in Table 3.1. We present the probability of correctly identify each MTD_k. The true dose-response curves for $K = 2, 3, 4$ represent a subset of the dose-response curves used in the corresponding scenario for $K = 5$ in Section 3.5. We note that population indices may change across $K = \{2, 3, 4\}$ but the $K = 5$ population index is reported within each scenario and K. For comparison, we simulated an independence design with 9 patients per population for each scenario and K. This design allowed dose-escalation after each patient under the restriction that no untried dose-levels be skipped when escalating, which we reiterate represents a high bar because this design would typically use cohorts of three patients (and a larger sample size), in practice.

We see that, in general, the probability of correctly identifying each MTD_k increases with K. Furthermore, we see that there is a clear advantage to HM with $K = 4$ but that the probability of correctly identifying each MTD_k was lower with the HM design than with the independence design with $K = 2$ due to the increased complexity of the hierarchical power model. With $K = 3$, HM increased the probability of correctly identifying each MTD_k, in most cases, but performed particularly poorly in Scenario 6, where there was substantial heterogeneity in the MTDs across populations. Based on these results, we recommend that HM in Phase I clinical trials be implemented with a minimum of 3 populations but 4 populations are likely required to fully realize the advantages of HM.

3.5 Discussion

We discussed HM for sharing information across populations in Phase I clinical trials. First, we presented hierarchical extensions to three, commonly used dose-toxicity models for Phase I oncology trials. These models allow for a different MTD in each population, while borrowing strength across populations, when appropriate, to achieve a more precise estimate of each population's MTD. We then proposed three DFGs for Phase I clinical trials using HM. The proposed DFGs allow us to take full advantage of HM while protecting patient safety by restricting dose escalation until it has been

Table 3.1: Results from 1000 simulated trials for the power model for different K. Below are the selection probabilities for the target dose. Results using HM are in bold, while results from an independent design are displayed in the next row. $K = 5$ group indices are listed for each Scenario within each K.

Scenario	Design	**K = 2**		**K = 3**			**K = 4**			
1		Group2		Group2			Group2			
	HM	**0.457**	**0.423**	**0.464**	**0.507**	**0.506**	**0.483**	**0.477**	**0.494**	**0.514**
	Ind.	0.466	0.449	0.454	0.469	0.442	0.452	0.462	0.461	0.469
2		Group2		Group2			Group2			
	HM	**0.29**	**0.436**	**0.298**	**0.409**	**0.495**	**0.326**	**0.363**	**0. 418**	**0.50**
	Ind.	0.244	0.461	0.24	0.398	0.454	0.256	0.295	0.365	0.461
3		Group1	Group5	Group1	Group3	Group5	Group1	Group3	Group4	Group5
	HM	**0.594**	**0.453**	**0.59**	**0.349**	**0.469**	**0.547**	**0.388**	**0.438**	**0.494**
	Ind.	0.621	0.459	0.499	0.341	0.462	0.603	0.343	0.422	0.469
4		Group1	Group4	Group1	Group2	Group4	Group1	Group2	Group3	Group4
	HM	**0.357**	**0.414**	**0.401**	**0.385**	**0.41**	**0.381**	**0.42**	**0.402**	**0.436**
	Ind.	0.368	0.365	0.345	0.329	0.383	0.354	0.343	0.341	0.37
5		Group1	Group4	Group1	Group2	Group4	Group1	Group2	Group4	Group5
	HM	**0.297**	**0.456**	**0.306**	**0.324**	**0.471**	**0.305**	**0.30**	**0.494**	**0.453**
	Ind.	0.255	0.461	0.271	0.301	0.43	0.276	0.332	0.451	0.442
6		Group3	Group5	Group1	Group3	Group5	Group1	Group2	Group3	Group5
	HM	**0.338**	**0.47**	**0.509**	**0.357**	**0.46**	**0.529**	**0.375**	**0.389**	**0.423**
	Ind.	0.358	0.633	0.58	0.327	0.601	0.604	0.387	0.341	0.617

shown that the current dose-level has an acceptable toxicity profile. Our simulation results suggest that all three models are able to borrow strength when the MTDs are constant across populations, resulting in a more precise estimate of each population's MTD, but the hierarchical power model is more robust when the populations are more heterogeneous. In addition, we found that the "321" DFG provided the best trade-off for estimating each population's MTD while protecting patient safety of the three DFGs considered. Finally, our simulation results suggest that HM would be beneficial with as few as three populations but independent designs are more effective with only two populations.

Returning to our motivating example of completing multiple, independent Phase I trials to evaluate a single agent in multiple populations with different background standards-of-care, our results are clear: completing independent designs for each population is not the optimal approach and parallel designs while using HM to share information across populations is more efficient. Our results indicate that in most cases the HM approach results in an increased probability of correctly identifying each population's MTD and an increased number of patients treated at each population's MTD while decreasing the percent of patients experiencing DLTs and the number of patients treated at unsafe doses above each population's MTD. This provides a strong theoretical basis for pursuing this type of design but additional work is needed to identify the practical challenges related to implementing this approach.

We have presented the results of this chapter as an extension of the CRM, but the method discussed in Section 3.2 can be applied to Bayesian adaptive Phase I designs more broadly. Some clinicians remain hesitant to implement the CRM due to concerns about escalating too quickly, resulting in excess DLTs. We proposed three DFGs to protect patient safely, with the "321" algorithm exhibiting the best performance, and our results indicate that HM actually results in fewer DLTs than independent CRM designs. Nevertheless, other modifications to protect patient safety could also be considered. For example, Faries (1994) suggests always choosing the highest dose below the current estimate of the MTD. Alternately, the EWOC (escalation with overdose control) was proposed as an approach to limit DLTs in Phase I clinical trials (Babb

et al., 1998), while Neuenschwander et al. (2008) propose classifying the posterior probability of a DLT into four categories: under-dosing, targeted toxicity, excessive toxicity, and unacceptable toxicity, and using these probabilities to guide dose-finding. The aforementioned approaches represent changes to the dose-finding algorithm and not the dose-toxicity model. As a result, these methods could easily be integrated with HM to achieve the benefits of borrowing information across populations while further limiting DLTs.

Our simulation results presented in Section 3.4 depend on the prior distributions and prior dose-response skeletons discussed in Section 3.2. We considered a variety of prior input values and different prior distributions for our variance parameters, however, we found the general trends to be consistent. Increasing the domain for our smoothing parameter resulted in substantially improved performance when the populations are homogeneous but much worse performance when the populations are heterogeneous. We calibrated the hyperparameters for the hierarchical logistic regression and curve-free method to be consistent with the prior skeleton used in the hierarachical power model. While the hierarchical power model proved to be superior in this case, other skeletons might be more appropriate for the other two models. However, we expect that the non-ideal pooling behavior observed in the hierarchical logistic regression and curve-free models would be similar regardless of the skeleton. Furthermore, the standard power model is often used with the CRM and we believe that practitioners might be more familiar with specifying the power model skeleton than the corresponding parameters in the other models, which also supports the use of the hierarchical power model.

Phase I dose-escalation designs, like the CRM, attempt to identify the MTD under the assumption that the maximum dose that can be given safely is also the best choice for identifying an efficacious dose. In practice, it is often the case that the efficacy of a treatment may plateau or even diminish as dose increases, while potential toxicity is expected to increase monotonically with dose. This motivates the use of Phase I-II designs that consider both efficacy and toxicity during dose-finding. These designs often rely on a parametric model for the dose-response relationship for efficacy and toxicity and

it would be worthwhile to investigate whether HM would be beneficial in this scenario, as well.

3.6 Supplementary Materials

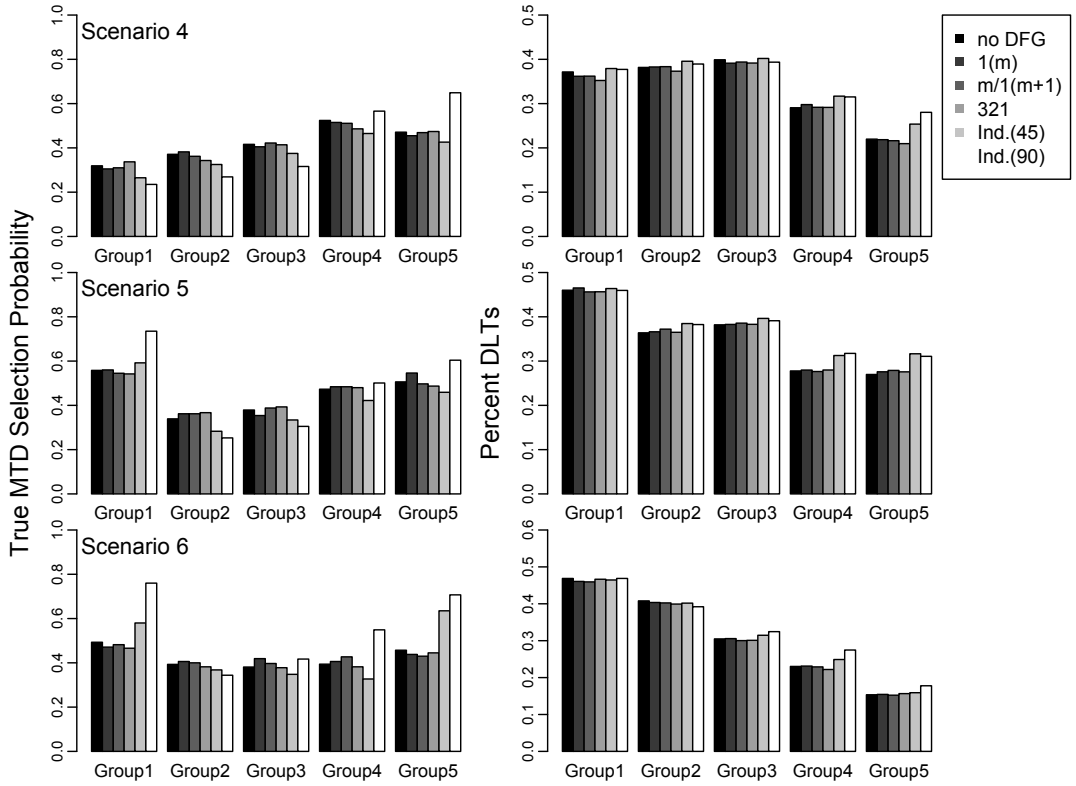

Figure 3.9: **Power Model**: Scenarios 4-6(left plots): Probability of correctly identifying the true MTD; Scenario 4-6(right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis.

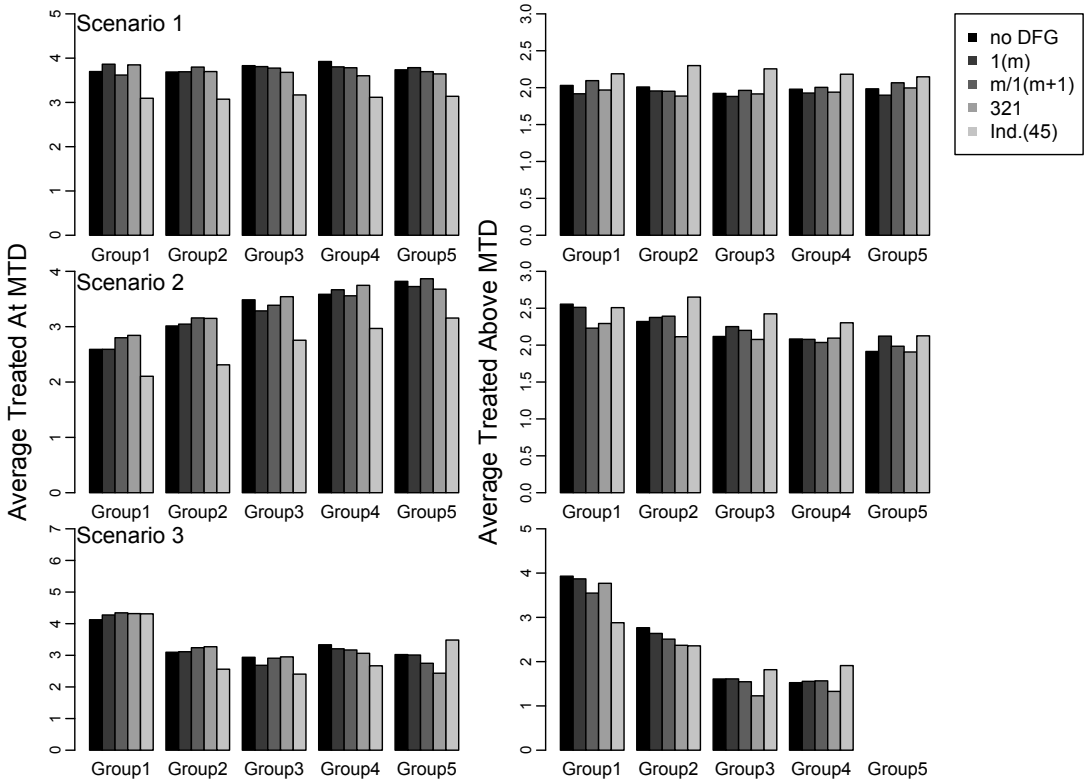

Figure 3.10: **Power Model**: Scenarios 1-3(left plots): Average number of patients treated at the true MTD; Scenario 1-3(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis.

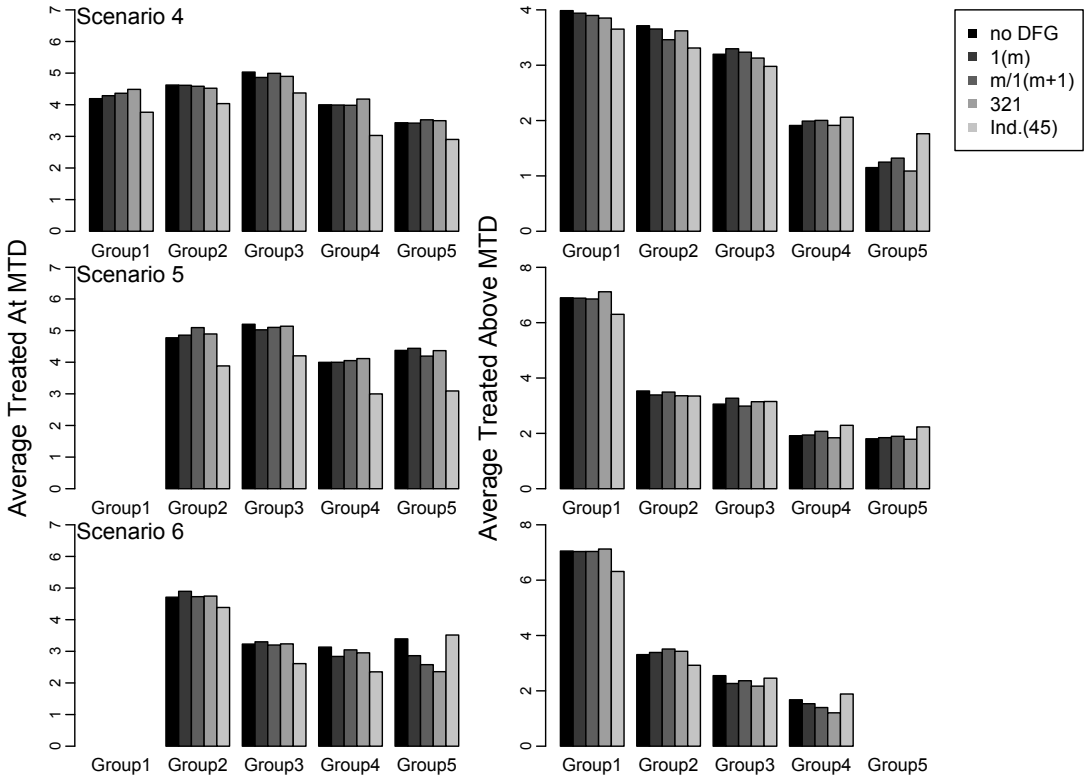

Figure 3.11: **Power Model**: Scenarios 4-6(left plots): Average number of patients treated at the true MTD; Scenario 4-6(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis.

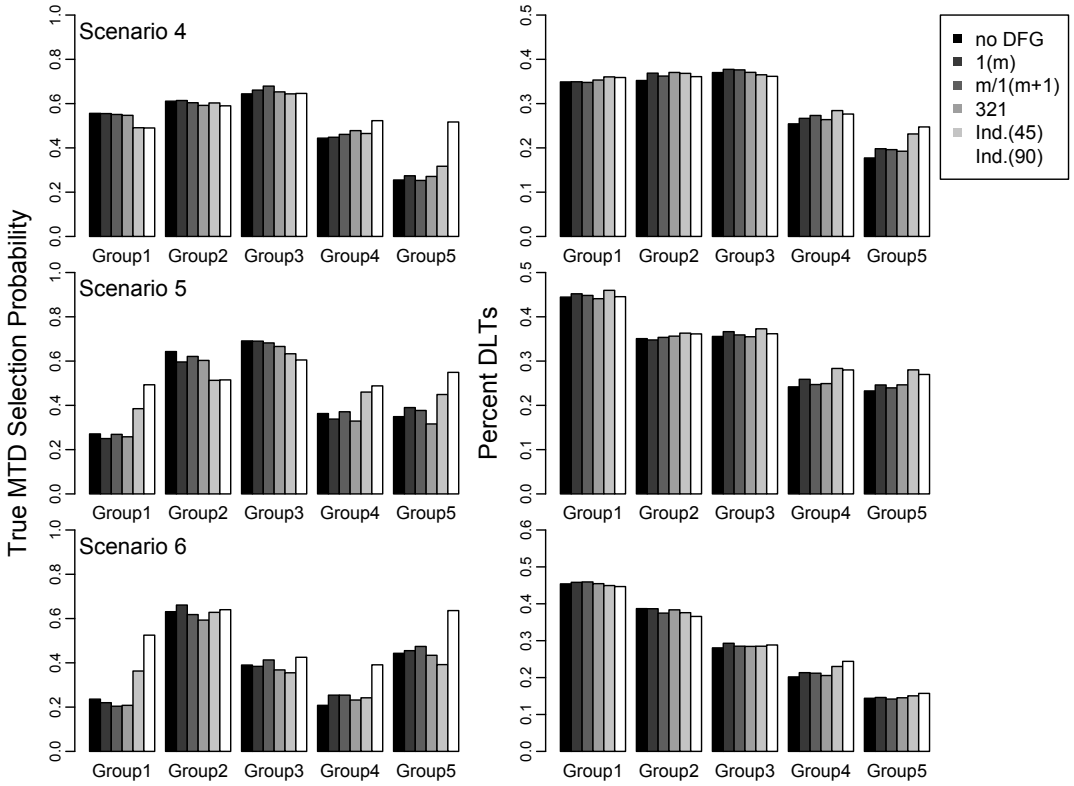

Figure 3.12: **Logistic Regression Model**: Scenarios 4-6(left plots): Probability of correctly identifying the true MTD; Scenario 4-6(right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis.

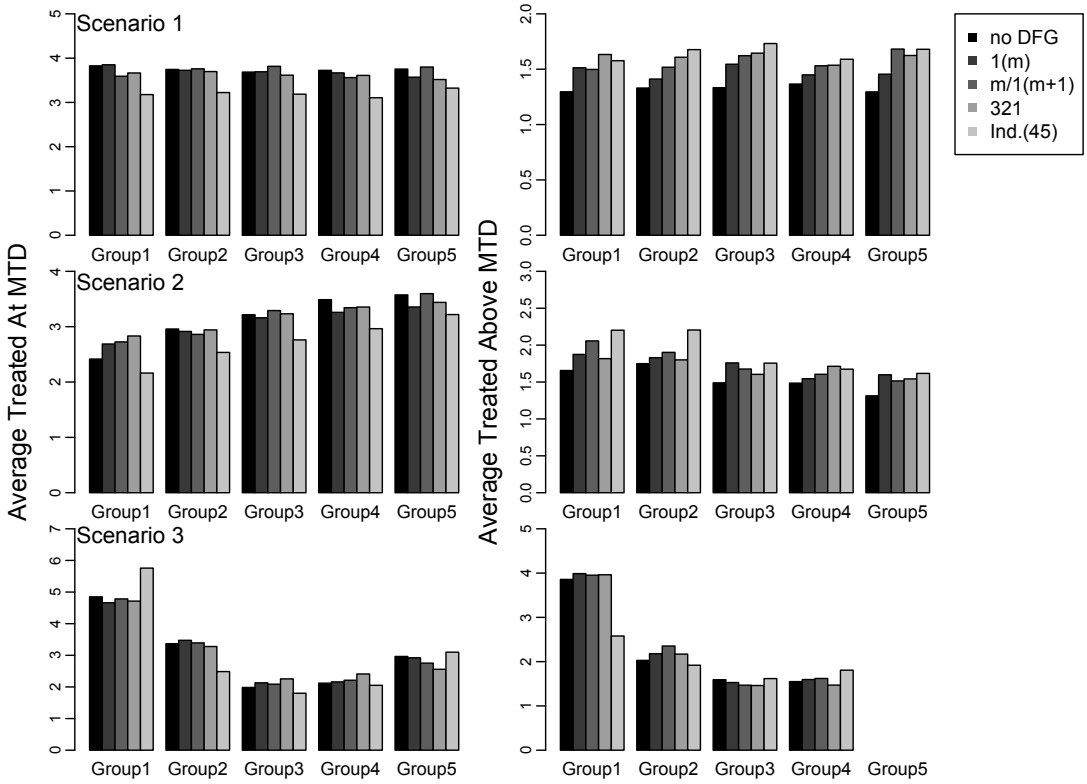

Figure 3.13: **Logistic Regression Model**: Scenarios 1-3(left plots): Average number of patients treated at the true MTD; Scenario 1-3(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis.

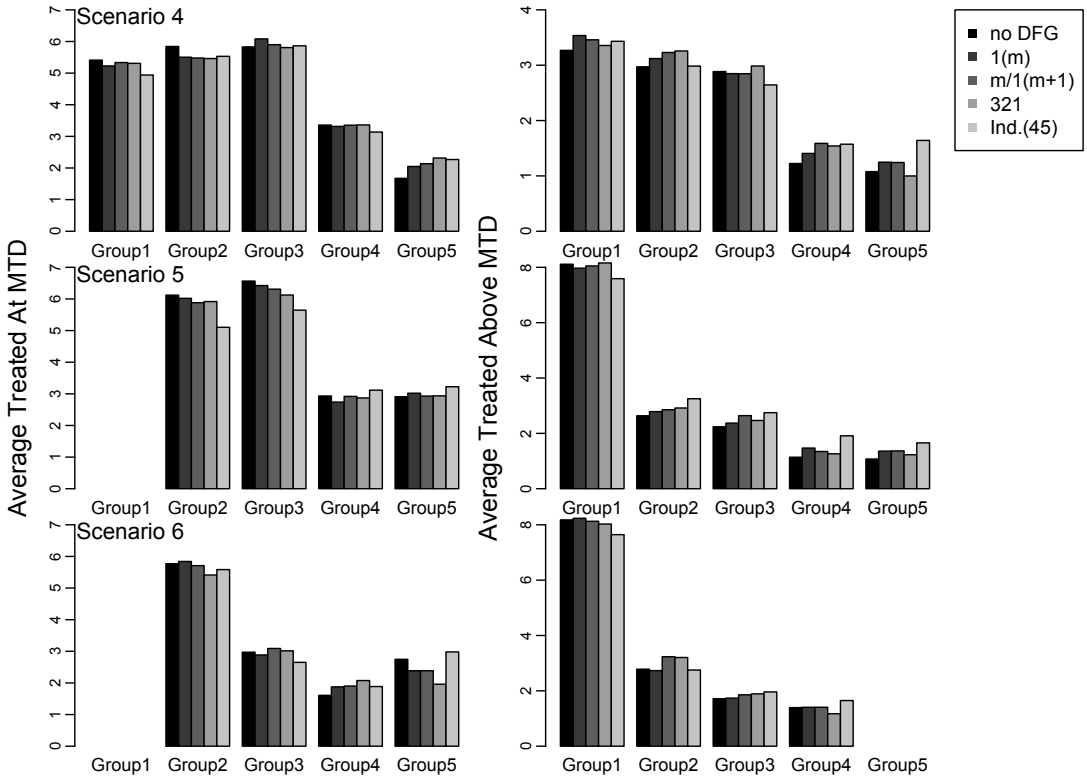

Figure 3.14: **Logistic Regression Model**: Scenarios 4-6(left plots): Average number of patients treated at the true MTD; Scenario 4-6(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis.

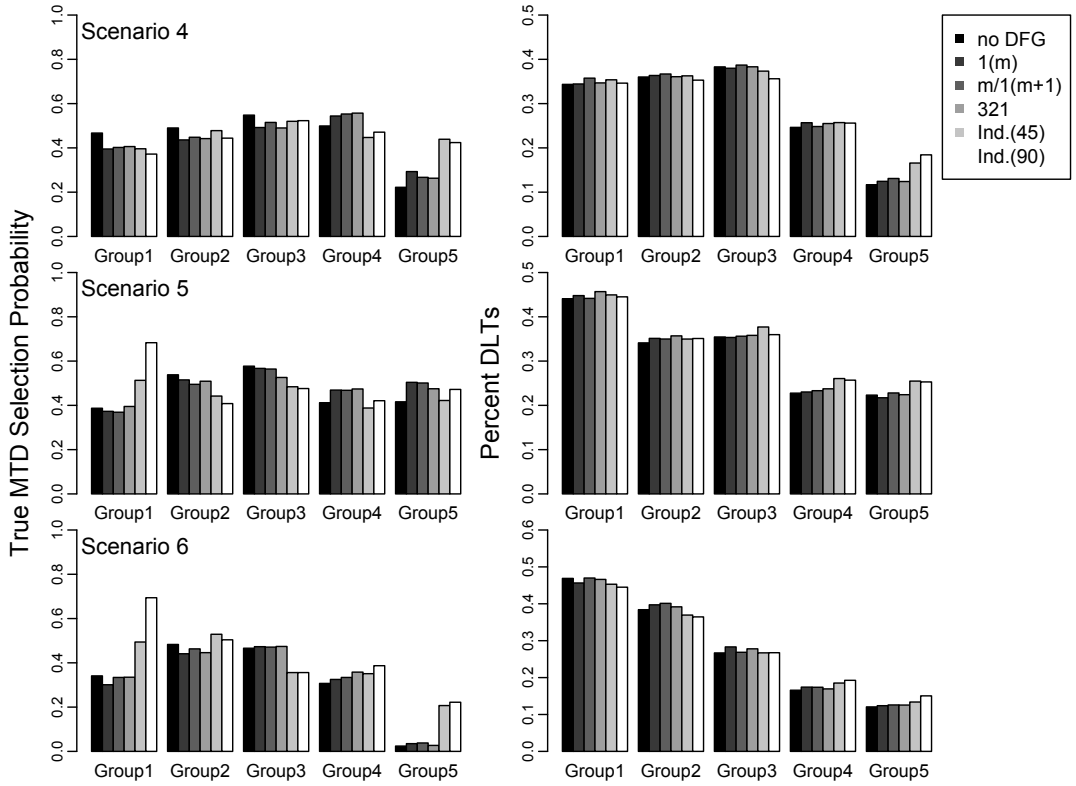

Figure 3.15: **Curve Free Model**: Scenarios 4-6(left plots): Probability of correctly identifying the true MTD; Scenario 4-6(right plots): The percent of patients experiencing a DLT for each population. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parenthesis.

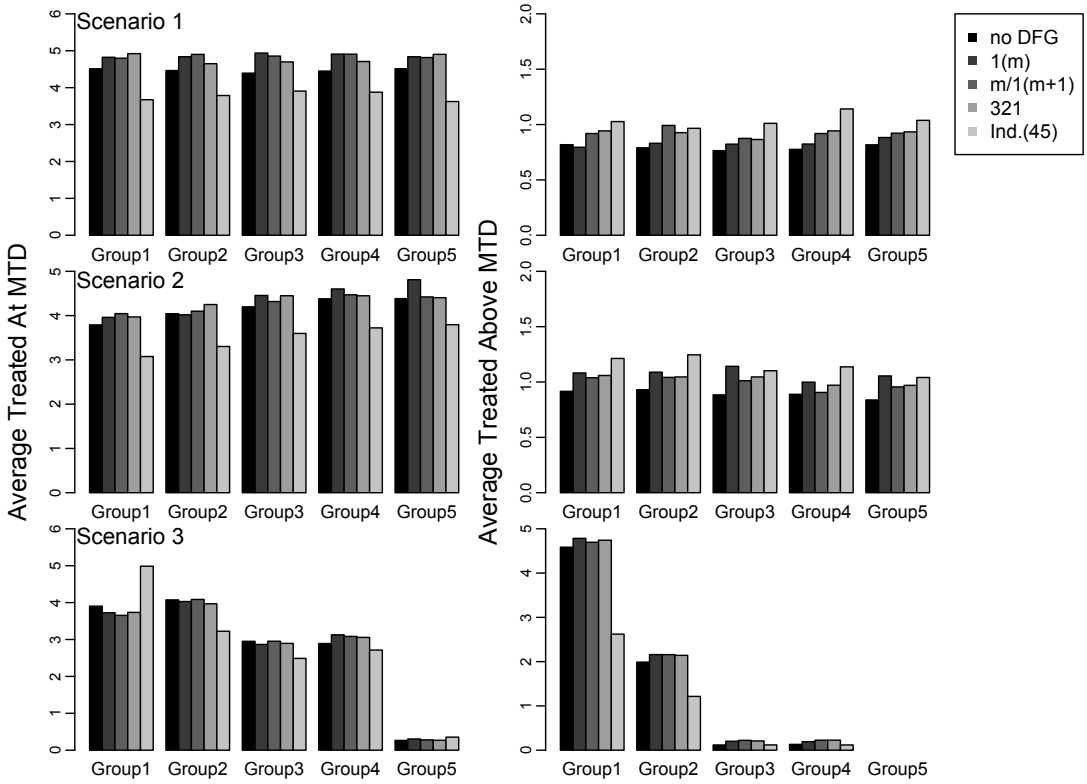

Figure 3.16: **Curve Free Model**: Scenarios 1-3(left plots): Average number of patients treated at the true MTD; Scenario 1-3(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis.

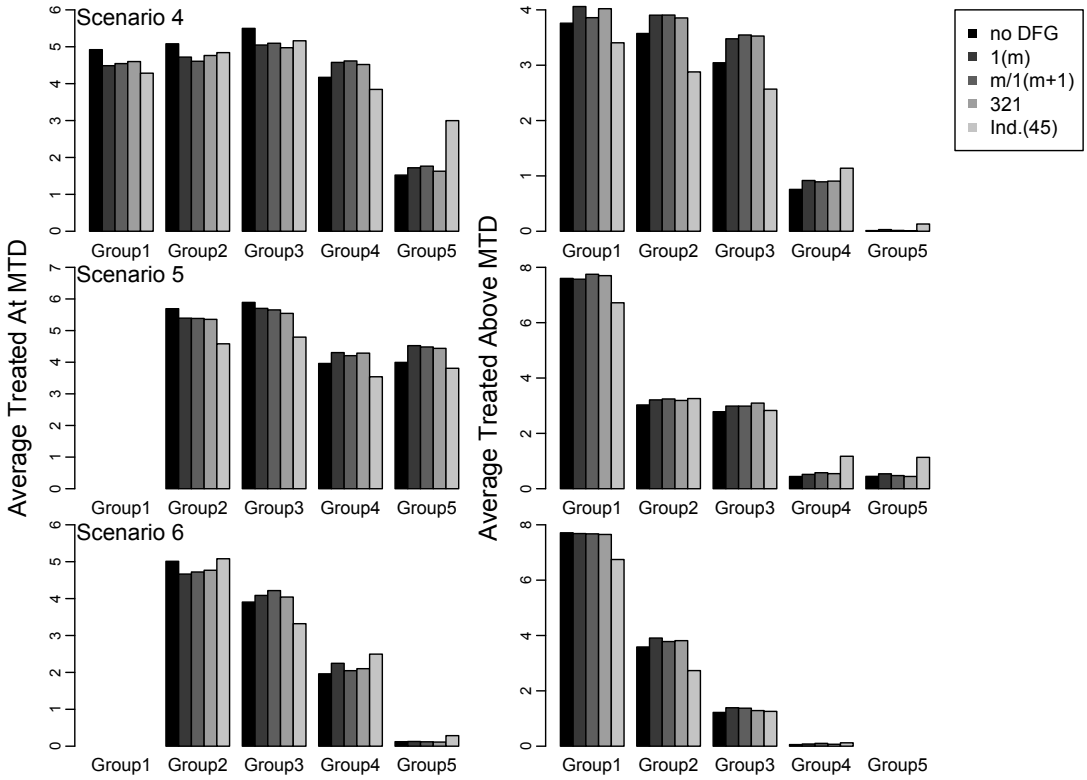

Figure 3.17: **Curve Free Model**: Scenarios 4-6(left plots): Average number of patients treated at the true MTD; Scenario 4-6(right plots): Average number of patients treated above the true MTD. Results from 1000 simulated trials are displayed for "no DFG" and the three proposed DFGs: "1(m)", "m/1(m+1)" and "321" (with max 45 patients), and K independent model assuming a maximum sample sizes of 45 patients, displayed in parenthesis.

Chapter 4

Efficacy/Toxicity Dose-Finding Using Hierarchical Modeling for Multiple Populations

4.1 Introduction

Phase I oncology trials are primarily dose-escalation studies to evaluate the safety of a novel treatment and identify the maximum tolerable dose (MTD), defined as the highest dose with probability of dose limiting toxicity (DLT) less than some pre-specified threshold. Typically, the efficacy of the new treatment is not examined until Phase II. The rationale for this approach is that clinicians have historically believed the probabilities of toxicity and treatment efficacy increase monotonically with dose and, as a result, the highest dose with acceptable toxicity was thought to have the best chance to succeed in future Phase II and III clinical trials. However, for contemporary biologically targeted agents, investigators often believe a drug's potential efficacy may level off or even diminish before reaching the MTD, while potential toxicity continues to increase with dose. This motivates dose-finding trials based on the simultaneous evaluation of toxicity and efficacy. Furthermore, given the limited sample sizes in Phase I oncology trials, incorporating efficacy into dose-finding may help to better identify the optimal dose used in subsequent Phase II and III clinical trials. In response, numerous Phase I-II

trial designs have been proposed that incorporate efficacy and toxicity into dose finding.

Gooley et al. (1994) were among the first to propose a dose-finding design based on simultaneous evaluation of toxicity and efficacy. Their results suggest that the additional dose-response curve for efficacy adds complexity to the dose selection algorithm. This suggests a need to consider the cost of additional model parameters (i.e., the response probability) when designing a Phase I-II trial. Consequently, Thall and Russell (1998) propose a design that combines toxicity and efficacy into one variable, thereby reducing the parameter space. Alternatively, Braun (2002) extends the continual reassessment method to account for two competing outcomes, while Thall and Cook (2004) take a similar approach but also define an efficacy/toxicity trade-off contour that can be used to guide dose-finding. A number of extensions to this basic approach have been discussed in the literature over the last decade (Yin et al., 2006; Zhang et al., 2006; O'Quigley et al., 2001; Houede et al., 2010; Ivanova, 2003; Nebiyou Bekele and Shen, 2005; Thall et al., 2008; Yuan and Yin, 2008; Thall et al., 2013; Koopmeiners and Modiano, 2014).

Researchers are often interested in evaluating the performance of a novel treatment in a number of patient populations, which may or may not have different background standards-of-care. Typically, investigators complete separate dose-escalation studies to establish the MTD in each population, which is an expensive and time-consuming process. Furthermore, while it is important to understand the behavior of the new drug in each population, it is also likely that the performance will be similar across populations, in which case researchers could gain efficiency and more precisely identify the optimal dose by sharing information across populations. This motivates a hierarchical modeling (HM) approach where parallel designs are run in each patient population but information is shared across populations to gain efficiency.

In Chapter 3, we investigated HM in the context of Phase I dose-escalation studies. We proposed three hierarchical extensions of commonly used dose-toxicity models in Phase I clinical trials and proposed dose-finding guidelines that protect patient safety, while allowing the design to fully realize the potential of HM. Our simulation results indicate incorporating HM into Phase I dose-finding trials results in an increase in

the probability of correctly identifying the MTD and the average number of patients treated at the MTD, with little impact on the rate of DLTs. In this chapter, we propose a Bayesian adaptive Phase I-II dose-escalation design that uses HM to estimate population-specific optimal doses, while sharing both dose-toxicity and dose-efficacy information across populations. First, we discuss three hierarchical extensions to commonly used probability models for efficacy and toxicity in Phase I-II clinical trials and adapt the dose-finding algorithm proposed in Chapter 3 to Phase I-II clinical trials. Our simulation results indicate that HM results in an increased probability of correctly identifying the optimal dose and increases the average number of patients treated at the optimal dose, with limited impact on the percent of DLTs observed in the trial.

The remainder of this chapter proceeds as follows. In Section 4.2, we describe three joint probability models for toxicity and efficacy that incorporate HM for sharing information across populations in Phase I-II dose-finding trials. First, we consider both parametric and non-parametric bivariate binary outcome model, and, in addition, we consider an under-parameterized model that combines toxicity and efficacy into a single trinary outcome. In Section 4.3, we describe our dose-finding algorithm and present simulation results evaluating the operating characteristics of our proposed design in Section 4.4. Finally, we conclude with a discussion in Section 4.5.

4.2 Models

In this section, we present hierarchical extensions of three joint probability models for efficacy and toxicity that have been proposed for use in Phase I-II dose-finding trials. In each case, we define a two-level Bayesian hierarchical model where the first level specifies the population-level parameters and the second level facilitates borrowing across populations. Existing joint probability models for Phase I-II clinical trials can be broadly classified into two groups: bivariate outcome models, where separate dose-response models are specified for efficacy and toxicity and the correlation between efficacy and toxicity is incorporated into the model using a copula model or some other approach (Braun, 2002; Thall and Cook, 2004; Yin et al., 2006), and trinomial models, where efficacy and toxicity are combined into a trinomial outcome and a dose-response

relationship is specified for the trinomial outcome (Thall and Russell, 1998; Zhang et al., 2006). We begin by discussing hierarchical extensions of two bivariate binary outcome models and then discuss a hierarchical extension of the trinomial model proposed by Zhang et al. (2006).

4.2.1 Bivariate Binary Outcomes

We use the following notation throughout Section 4.2.1. First, let T_{ikj} be a binary indicator for the presence or absence of DLT in subject i treated at dose j in population k, which takes the value 1 with probability $\pi_{T,kj}$, and let E_{ikj} be a binary indicator for the probability of tumor response in subject i treated at dose j in population k, which takes the value 1 with probability $\pi_{E,kj}$. We will consider two approaches for specifying a bivariate outcome model. First, we consider a parametric approach, where parametric dose-response models are specified for efficacy and toxicity. Next, we consider a non-parametric model that imposes a monotonicity constraint on the dose-toxicity model but avoids a formal parametric model.

Parametric Model

For our parametric model, we extend a simple one-parameter power model for toxicity and a more flexible, quadratic logistic regression model for efficacy. Our hierarchical model for toxicity is specified as:

$$pr(T_{ikj} = 1|population = k, dose = j) = \pi_{T,kj} = p_j^{exp(\alpha_k)} \tag{4.1}$$

$$\alpha_k|\mu_T, \sigma_T^2 \sim N(\mu_T, \sigma_T^2)$$

$$\mu_T \sim Normal(0, 2^2) \quad \text{and} \quad \sigma_T \sim Uniform(0.39, 3)$$

for dose level $j = 1, \ldots, D$ and population $k = 1, \ldots, K$. The vector $(p_1, \ldots, p_D)$ is referred to as the skeleton and its components are monotonically increasing and take values between 0 and 1. For our simulation results presented in Section 4.4, we set the power model skeleton equal to $(0.05, 0.15, 0.25, 0.35, 0.45)$. Our hierarchical model for efficacy is specified as:

$$pr(E_{ikj} = 1|population = k, dose = j) = \pi_{E,kj} = \beta_{0k} + \beta_{1k}(dose - 1) + \beta_{2k}(dose - 1)^2, \tag{4.2}$$

$$\beta_{lk}|\mu_l, \sigma_l^2 \sim Normal(\mu_l, \sigma_l^2)$$
$$\mu_l \sim Normal(m_l, s_l^2) \quad \text{and} \quad \sigma_l \sim Uniform(0.39, 3),$$

for $l = 0, 1, 2$, dose level $j = 1, \ldots, D$ and population $k = 1, \ldots, K$. We originally fixed the intercept equal to -3 to reduce the number of unknown parameters, as suggested by Goodman et al. (1995). This reflects a 5% probability of tumor response at dose level 1, but we found that this model did not provide enough flexibility when the true optimal dose resides in the higher dose levels. The unknown m_0, m_1, and m_2 are the shared mean hyper-parameters for the intercept, linear and quadratic terms and are set equal to -2, 0.1, and 0, respectively, with shared variance hyper-parameters set to $s_0^2 = 4$, $s_1^2 = 9$, and $s_2^2 = 4$. This corresponds to a conservative, monotonic prior efficacy-skeleton of 0.12, 0.13, 0.14, 0.15, 0.17 for dose levels 1, 2, 3, 4, 5, respectively. The σ_l^2 are our hierarchical variance parameters that control the amount of borrowing across populations, with smaller values indicating more borrowing. We specify a uniform prior distribution on the standard deviation, rather than the log standard deviation, as in Chapter 3, since this prior is well-received for other hierarchical applications and we are interested in exploring its use further in our dose-finding setting. In Chapter 3, a uniform prior on the standard deviation with a lower bound of 0 produced poor convergence and identifiability, given the small sample sizes early in a trial. The lower bound of our uniform prior was set to 0.39, based on our simulation results, which suggested that a lower bound less than 0.39 results in over-borrowing and poor trial operating characteristics.

The toxicity and efficacy outcomes in Phase I-II clinical trials are thought to be correlated and previous authors proposing bivariate outcome models for Phase I-II clinical trials have used copula models to specify the correlation (Thall and Cook, 2004; Braun, 2002). Copula models are complex models that require a large amount of data to properly model the correlation between outcomes. Previously, we illustrated that the sample sizes found in Phase I-II clinical trials are typically inadequate to estimate the correlation parameters in copula models and that the performance of Phase I-II trial designs is not diminished if the two endpoints are assumed to be independent, even in the presence of strong correlation (Cunanan and Koopmeiners, 2014). Other authors have reported similar results (Yin et al., 2006). Therefore, we will proceed assuming independence

between the toxicity and efficacy outcome for our parametric model.

Non-Parametric Model

The second model we consider is a hierarchical extension of the non-parametric model proposed by Yin et al. (2006). They specify a dose-response relationship for toxicity and efficacy through the following transformations. For population $k = 1, \ldots, K$, the dose-response model for toxicity is specified as,

$$\phi_{k1} = logit(\pi_{T,k1}), \quad \phi_{kj} = log\left(\frac{\pi_{T,kj}}{1-\pi_{T,kj}} - \frac{\pi_{T,k(j-1)}}{1-\pi_{T,k(j-1)}}\right),$$

for $j = 2, \ldots, D$, and for efficacy, let

$$\psi_{k1} = logit(\pi_{E,k1}), \quad \psi_{kj} = log\left(\frac{\pi_{E,kj}}{1-\pi_{E,kj}}\right) - log\left(\frac{\pi_{E,k(j-1)}}{1-\pi_{E,k(j-1)}}\right),$$

for $j = 2, \ldots, D$. The primary difference between the two parameterizations is that the model for toxicity enforces a monotonicity constraint on the dose-response relationship for toxicity, whereas the model for efficacy does not. Yin et al. (2006) originally specified a bivariate normal prior for the efficacy and toxicity parameters to allow a prior correlation between the model parameters but found that setting the off-diagonal covariance elements to zero did not impact their results. We will specify independent normal priors for $\phi_{j,k}$ and $\psi_{j,k}$ and facilitate borrowing strength across populations by specifying a hierarchical model on ϕ_{kj} and ψ_{kj} as follows:

$$\phi_{kj}|\mu_{\phi,j}, \sigma^2_{\phi,j} \sim Normal(\mu_{\phi,j}, \sigma^2_{\phi,j}) \tag{4.3}$$

$$\mu_{\phi,j} \sim Normal(0, 50) \quad \text{and} \quad \sigma_{\phi,j} \sim Uniform(0.39, 3)$$

and,

$$\psi_{kj}|\mu_{\psi,j}, \sigma^2_{\psi,j} \sim Normal(\mu_{\psi,j}, \sigma^2_{\psi,j}) \tag{4.4}$$

$$\mu_{\psi,j} \sim Normal(0, 50) \quad \text{and} \quad \sigma_{\psi,j} \sim Uniform(0.39, 3)$$

for dose level $j = 1, \ldots, D$ and population $k = 1, \ldots, K$. Yin et al. (2006) specify a $Normal(0, 100)$ prior on ϕ_j and ψ_j. Our design, which uses HM to share information across populations, will have a smaller sample size for each population than would typically be used in an independent Phase I-II design. To accommodate the smaller sample

size, we reduce the prior variance to 50. Similar to the parametric binary bivariate model, $\sigma^2_{\psi,j}$ and $\sigma^2_{\psi,j}$ are our hierarchical variance parameters, controlling the amount of sharing across populations, and were specified using simulation studies, as described at the end of this section.

The dose-response models specified above provide the marginal probabilities of toxicity and efficacy. Yin et al. (2006) induce correlation between the toxicity and efficacy outcomes using the global cross-ratio model proposed by Dale (1986). Define $\pi_{xy,kj} = Pr(T_{kj} = x, E_{kj} = y|\text{population} = k, \text{dose} = j)$ with $x \in \{0, 1\}$ and $y \in \{0, 1\}$. Under the global cross-ratio model, the toxicity-efficacy odds ratio (OR) is defined as follows:

$$\theta_{kj} = \frac{\pi_{00,kj}\pi_{11,kj}}{\pi_{01,kj}\pi_{10,kj}},$$

where θ_{kj} quantifies the association between the two outcomes for population k at dose level j. Yin et al. (2006) specify a $logNormal(0, 10)$ prior distribution for each θ_j and assume all θ_j's are independent to ease computation. To reduce our parameter space, we define $\theta_{kj} \sim logNormal(0, 5)$, rather than define a hierarchical structure to share information across populations when estimating the odds ratio. Recall that we will share information across populations for estimating the probability of toxicity and efficacy through the hierarchical models specified in Equations (4.3) and (4.4) and feel that specifying an additional hierarchy for the odds ratio is unnecessary. Finally, we reduce the prior variance for the odds ratio to accommodate a smaller sample size for each population compared to an independent design for each population. After accounting for the correlation induced by the global cross-ratio model, the joint toxicity and efficacy outcomes for dose j and population k follow a multinomial distribution with response probabilities $(\pi_{11,kj}, \pi_{10,kj}, \pi_{01,kj}, \pi_{00,kj})$ and a sample size of n_{kj} patients, where the

response probabilities are defined as follows (Dale, 1986):

$$\pi_{11,kj} = \begin{cases} (a_{kj} - \sqrt{a_{kj}^2 + b_{kj}}/\{2(\theta_{kj} - 1)\}, & \text{for } \theta_{kj} \neq 1 \\ \pi_{T,kj}\pi_{E,kj}, & \text{for } \theta_{kj} = 1 \end{cases}$$

$$\pi_{10,kj} = \pi_{T,kj} - \pi_{11,kj}$$

$$\pi_{01,kj} = \pi_{E,kj} - \pi_{11,kj}$$

$$\pi_{00,kj} = 1 - \pi_{T,kj} - \pi_{E,kj} + \pi_{11,kj},$$

with $a_{kj} = 1 + (\pi_{T,kj} + \pi_{E,kj})(\theta_{kj} - 1)$ and $b_{kj} = (-4)\theta_{kj}(\theta_{kj} - 1)\pi_{T,kj}\pi_{E,kj}$.

4.2.2 Trinary Outcome

The last model we consider is a hierarchical extension of the triCRM proposed by Zhang et al. (2006). Rather than separately modeling bivariate binary outcomes, they collapse the four possible outcomes into a single variable with three outcomes: no efficacy or toxicity, efficacy without toxicity and toxicity with or without efficacy. An advantage to this approach is that the model is simple relative to the other models, since we do not have to model separate dose-response models for the two outcomes, but a disadvantage is that the marginal probability of efficacy is no longer identifiable. However, our primary interest is identifying doses with sufficient efficacy and acceptable toxicity, mitigating the impact of this disadvantage.

Denote the probabilities of the three possible outcomes (no efficacy or toxicity, efficacy without toxicity, toxicity with or without efficacy) as ψ_0, ψ_1, ψ_2, respectively, which by definition sum to 1. We can define a hierarchical extension of the continuation-ratio model proposed by Zhang et al. (2006) as follows:

$$log\left(\frac{\psi_{1,kj}}{\psi_{0,kj}}\middle| population = k, dose = j\right) = \alpha_{1k} + \alpha_{2k} + \gamma_{1k}(dose) \tag{4.5}$$

$$logit(\psi_{2,kj}|population = k, dose = j) = \alpha_{1k} + \gamma_{2k}(dose),$$

for dose levels $j = 1, \ldots, D$ and population $k = 1, \ldots, K$, with hierarchical priors defined as follows:

$$\begin{aligned}
\alpha_{tk}|\mu_{\alpha t}, \sigma^2_{\alpha t} &\sim Normal(\mu_{\alpha t}, \sigma^2_{\alpha t}) \\
\gamma_{tk}|\mu_{\gamma t}, \sigma^2_{\gamma t} &\sim Normal(\mu_{\gamma t}, \sigma^2_{\gamma t}) \\
\mu_{\alpha t} \sim Normal(u_t, c_t^2) \quad &\text{and} \quad \sigma_{\alpha t} \sim Uniform(0.39, 3) \\
\mu_{\gamma t} \sim Normal(v_t, b_t^2) \quad &\text{and} \quad \sigma_{\gamma t} \sim Uniform(0.39, 3)
\end{aligned}$$

for $t = 1, 2$ with $u_1 = -3, u_2 = 2, v_1 = 0.5, v_2 = 1$, and $c_1 = c_2 = b_1 = b_2 = 4$. The second level mean specifications correspond to a prior toxicity skeleton, i.e., ψ_2, of 0.12, 0.27, 0.50, 0.73, 0.88 for dose levels 1, 2, 3, 4, 5, respectively; and a prior skeleton for efficacy with no toxicity, i.e., ψ_1, of 0.33, 0.37, 0.31, 0.20, 0.10 for dose levels 1, 2, 3, 4, 5, respectively, which results in a prior skeleton for no response, i.e., ψ_0, of 0.55, 0.37, 0.19, 0.07, 0.02 for dose levels 1, 2, 3, 4, 5, respectively. As in the binary bivariate models, $\sigma^2_{\alpha t}$ and $\sigma^2_{\gamma t}$ are our hierarchical variance parameters, controlling the amount of sharing across populations, and were specified using simulation studies, as described in Section 4.2.3.

There is one major difference between our hierarchical model and the original model proposed by Zhang et al. (2006). For simplicity and computational ease, Zhang et al. (2006) define a $Uniform(-10, 5)$ prior for the common intercept α_{1k}, a $Uniform(0, 10)$ prior on the second intercept α_{2k}, and $Uniform(0, 10)$ priors on the slope parameters γ_{1k} and γ_{2k}. These priors impart the following restrictions on the model: (i) the probability of no response, ψ_0, decreases monotonically with dose, (ii) the probability of toxicity with or without efficacy, ψ_2, increases monotonically with dose, and (iii) the probability of efficacy without toxicity, ψ_1, may or may not be monotone with dose. In contrast, our hierarchical model has no such restrictions. We originally considered hierarchical prior specifications that maintained these restrictions but found them to be difficult to implement computationally. Furthermore, our simulation results suggest that our model performs well without these restrictions and, hence, we proceed with the hierarchical model presented above.

4.2.3 Hyperparameter Specification

The hyperparameters for the models discussed above were determined by simulation, as follows. The second-level mean hyperparameters were determined by separately varying each parameter in the corresponding independence model and selecting the combination with the most robust performance, as evaluated by simulation. We specify a uniform prior for the second-level standard deviation. The lower bound for the uniform prior distribution was set greater than zero due to the small sample sizes found in Phase I clinical trials, especially early in the trial, in which case the model cannot rule out a population variance of zero, resulting in an invalid distribution for our first level probability model. After fixing the mean hyperparameters, we selected the hyperparameters for the standard deviation for each model by progressively increasing the lower bound for our uniform prior and selecting the value with the most robust operating characteristics, as evaluated by simulation. We note simulations were similar for a larger lower bound for the non-parametric bivariate binary model, however, we chose the smaller value to be consistent with the other models.

4.3 Dose-Finding Algorithm

In this section, we discuss dose-finding when using hierarchical modeling to share information across populations in Phase I-II clinical trials. We expect enrollment to be staggered and randomly distributed across populations. In Chapter 3, we discussed three dose-finding guidelines that define when to allow dose-escalation within a population taking into account the number of patients observed in other populations. These guidelines are incorporated for patient safety but our simulation results suggest that our guidelines result in improved operating characteristics based on a number of metrics, compared to unrestricted dose-finding. In this chapter, we consider only a single dose-finding guideline based on our results from Chapter 3.

We identify a set of acceptable doses for each population, assuming admissibility criteria and minimum performance levels as elicited from clinicians. For each population, we determine the optimal dose from the set of acceptable doses by maximizing each population's posterior mean probability of efficacy without toxicity, following the work

of Yin et al. (2006). For the parametric bivariate model, the two binary outcomes are assumed independent and the probability of efficacy with no toxicity for each dose is simply the product of the marginal probability of efficacy and the marginal probability of toxicity. For the non-parametric bivariate model, the optimal dose is determined by π_{01}, the multinomial probability for efficacy with no toxicity. For the two bivariate binary outcome models described in Section 4.2.1, a dose is acceptable if the posterior probability of a DLT being less than the clinician-specified target toxicity level and the posterior probability of an efficacious response exceeding the clinician-specified minimum threshold for efficacy both exceed pre-specified minimum thresholds, i.e.,

$$Pr(\pi_{T_k} < \overline{\pi}_T | Data, Dose) > \gamma_T \quad \text{and} \quad Pr(\pi_{E_k} > \underline{\pi}_E | Data, Dose) > \gamma_E \tag{4.6}$$

where $\overline{\pi}_T$ is the maximum acceptable probability of DLT, $\underline{\pi}_E$ is the minimum acceptable probability of efficacy, and γ_T and γ_E are the minimum pre-specified thresholds for toxicity and efficacy, respectively. These are admissibility criteria for toxicity and efficacy proposed by Thall and Cook (2004). The thresholds γ_T and γ_E are typically chosen between 0.05 and 0.20 and can be thought of as tuning parameters to achieve desired trial operating characteristics (Berry et al., 2010).

We cannot use the acceptability criteria described above for the trinary model because although the marginal probability of toxicity can be estimated, the marginal probability of efficacy is not identifiable. Instead, we use two decision functions proposed by Zhang et al. (2006) to determine the set of acceptable doses and, among those found to be acceptable, the optimal dose. We denote $\hat{\psi}_{0,kj}$, $\hat{\psi}_{1,kj}$, $\hat{\psi}_{2,kj}$ to be the posterior mean probabilities of no toxicity or efficacy, efficacy without toxicity, and toxicity with or without efficacy, respectively. The first decision rule determines the set of acceptably safe doses using:

$$\delta_{1,kj} = I(\hat{\psi}_{2,kj} < \overline{\pi}_T) \tag{4.7}$$

Given that a dose is acceptable, i.e., $\delta_{1,kj} = 1$, Yin et al. (2006) determine the optimal dose from the set of acceptable doses by maximizing the toxicity-adjusted treatment success rate,

$$\delta_{2,kj} = \hat{\psi}_{1,kj} - \lambda \hat{\psi}_{2,kj}, \tag{4.8}$$

where $0 \leq \lambda \leq 1$ is a weight for the posterior mean probability of toxicity, $\hat{\psi}_{2,kj}$. If we set $\lambda = 0$, the decision rule to determine the optimal dose is the dose maximizing the posterior mean probability of efficacy with no toxicity. We can also consider using $\delta_{1,kj}$ for the two models presented in Section 4.2.1. However, these models are over-specified and using $\delta_{1,kj}$ with these models results in more trials stopping early due to poor estimation. Finally, we also modify Zhang et al. (2006)'s decision function, $\delta_{1,kj}$, to require a minimum performance level for efficacy,

$$\delta^*_{1,kj} = I(\hat{\psi}_{2,kj} < \overline{\pi}_T) \times I\left(\frac{\hat{\psi}_{1,kj}}{\hat{\psi}_{0,kj} + \hat{\psi}_{1,kj}} > \underline{\pi}_{E|T^c}\right). \tag{4.9}$$

That is, we require the posterior mean probability of efficacy conditional on no toxicity to be greater than some minimum pre-specified threshold in addition to the toxicity decision criteria found in (4.7).

In a standard Phase I-II clinical trial, the initial cohort of (typically) three patients is treated at the lowest dose level and subsequent cohorts are treated at the current estimate of the optimal dose based on the outcomes for all previous subjects under the restriction that no untried dose-level may be skipped when escalating. Extending this approach to hierarchical modeling with multiple populations is not straightforward. One approach would be to escalate in cohorts of three patients regardless of the population, but this would be too aggressive and potentially result in a patient being treated at a dose-level before other patients from the same population are treated at a lower dose-level. Alternately, escalation could occur using cohorts of three patients within a population but this would not take full advantage of sharing information across populations using hierarchical modeling. Instead, we will use the "$m/1(m+1)$" dose finding guideline (DFG) described in Chapter 3. The "$m/1(m+1)$" DFG provides a run-in period for each population and indicates when a population is able to escalate to an untried dose, but the ultimate decision to escalate is based on the current estimate of the optimal dose. Formally, the "$m/1(m+1)$" DFG allows escalation to dose-level $j+1$ for population $k = 1, \ldots, K$ if:

- $\mathbf{m}$ patients in population k,
- Or $\mathbf{m+1}$ patients overall (and at least 1 patient in population k),

have been treated at dose level j, for $j = 1, \ldots, D-1$. This DFG will encourage escalation, when appropriate, but was shown in Chapter 3 to effectively limit the number of DLTs and patients treated at overly toxic dose-levels.

Yin et al. (2006) propose that the trial should escalate to the next untried dose level if there is high posterior probability that the probability of DLT for the highest tried dose is less than the target toxicity level, i.e.,

$$Pr(\pi_{T_k} < \overline{\pi}_T | Data, Dose_{max}) > p, \tag{4.10}$$

where $p \geq \gamma_T$. With the complex models used in Phase I-II clinical trials, it can be difficult to estimate the dose-response curves when there are multiple untried dose-levels. Criterion (4.10) encourages escalation when there are untried dose levels that appear to be sufficiently safe. Once a population is allowed to escalate, as determined by the "$m/1(m+1)$" DFG, we implement the above escalation rule.

In summary, our proposed Phase I-II design will proceed as follows:

1. Treat the first patient in each population at the lowest dose level.

2. When a new patient is enrolled, determine if their population is allowed to escalate (or de-escalate) as determined by the "$m/1(m+1)$" DFG and if so, update the posterior distribution for all model parameters using all available data. Otherwise, treat the next patient at the current dose-level.

3. If escalation is allowed, determine the set of acceptable dose-levels for the current population using Criteria (4.6) for the bivariate models or Criteria (4.9) for the trinomial model. The trial terminates for futility if all dose-levels are unacceptable and Criterion (4.10) is not satisfied.

4. Otherwise, treat the next patient at the dose level that maximizes either $\pi_{01,kj}$, for the bivariate models, or $\delta_{2,kj}$, for the trinomial model, under the restriction that untried dose-levels cannot be skipped when escalating. If Criterion (4.10) is satisfied, the next patient will be treated at the lowest untried dose-level.

5. Repeat steps 2-4 until the maximum overall sample size is reached. Within each population, the acceptable dose that maximizes either $\pi_{01,kj}$ or $\delta_{2,kj}$ (depending on the model) at study completion is considered the optimal dose.

4.4 Simulation Study

We conducted a simulation study to evaluate the operating characteristics of a Phase I-II clinical that incorporates hierarchical modeling using the models discussed in Section 4.2. Design performance was summarized by (i) the probability of correctly identifying each population's biologically optimal dose (BOD), (ii) the percent of patients experiencing a DLT, and (iii) the average number of patients treated at each population's BOD. Simulations were completed assuming the following design parameters. The target toxicity level was set to $\overline{\pi}_T = 0.3$ and the minimum efficacy level was set to $\underline{\pi}_E = 0.3$. Gatekeepers defining acceptable doses were set to $\gamma_T = 0.25$, $\gamma_E = 0.1$ for Criteria (4.6) and $\underline{\pi}_{E|T^c} = 0.1$ for Criteria (4.9). We set $p = 0.5$ for Criterion (4.10), which encourages escalation to untried doses if all tried doses have been shown to be safe. We assume $K = 5$ populations with uniform enrollment across populations and $D = 5$ dose levels for investigation. We assume a maximum sample size of 100 total patients, with a minimum of 3 patients per population. This corresponds to an average of 20 patients in each population. Gibbs and slice sampling were completed in JAGS via R using $rjags$ (Plummer, 2011). Posterior inference was completed using 5,000 MCMC samples following a period of 1,000 iterations for burn-in. 1000 simulated trials were completed for each scenario.

For comparison, we also evaluate the operating characteristics assuming independent trials were completed for each population. Simulations were completed with models analogous to the three models discussed in Section 4.2. In each case, we fit the model specified in Section 4.2 without the second level of the hierarchy. A maximum of 20 subjects per population was used for the independent designs.

4.4.1 Scenarios

Data were simulated from one of four scenarios. Each scenario included five populations. The dose-toxicity and dose-efficacy curves for each population were generated from one of the thirteen cases discussed by Yin et al. (2006). The dose-toxicity and dose-efficacy curves for each case are presented in Figure 4.1. We define the BOD to be the dose that maximizes π_{01} (black dot) or $\delta_2(\lambda = 0)$, with δ_1^* defining acceptable doses (square), depending on the model. Other decision rules for selecting the optimal dose include minimizing the toxicity-efficacy odds ratio (open circle) and maximizing $\delta_2(\lambda = 0)$ with Zhang et al. (2006)'s original decision function δ_1 to define acceptable doses (x). In Scenario 1, the same combination of dose-toxicity and dose-efficacy curves were used for all five populations (Case 1), where the lowest dose is the optimal dose. In Scenario 2, population one uses the curves from Case 3, where toxicity and efficacy increase at the same rate (OR = 1) and there is no optimal dose, and the other four populations use the curves from Case 13, where dose four is optimal. In Scenario 3, the populations are more heterogeneous and the optimal dose varies substantially across populations. Specifically, Scenario 3 is comprised of Cases 3, 9, 1, 8 and 11, which have no optimal dose, no optimal dose, optimal dose equal to dose one, optimal dose equal to dose three, and optimal dose equal to dose three, respectively. Finally, in Scenario 4, the optimal dose differs by population but is clustered around the intermediate dose-levels. Scenario 4 is comprised of Cases 6, 7, 8, 10 and 13, which have optimal doses of dose level four, three, three, two and four, respectively.

4.4.2 Results

Figure 4.2 shows results for Scenarios 1 and 2. Figure 4.2 presents the probability of correctly identifying the optimal dose (top plots), the percent of patients experiencing a DLT (middle plots), and the average number of patients treated at the optimal dose (bottom plots). In Scenario 1, all populations have the same dose-toxicity and dose-efficacy curves, resulting in the same optimal dose for all populations (dose level 1). We see that hierarchical modeling results in an increase in the probability of correctly identifying the optimal dose with the parametric bivariate model (displayed as "PLR HM") showing the best performance, while the non-parametric bivariate model showed

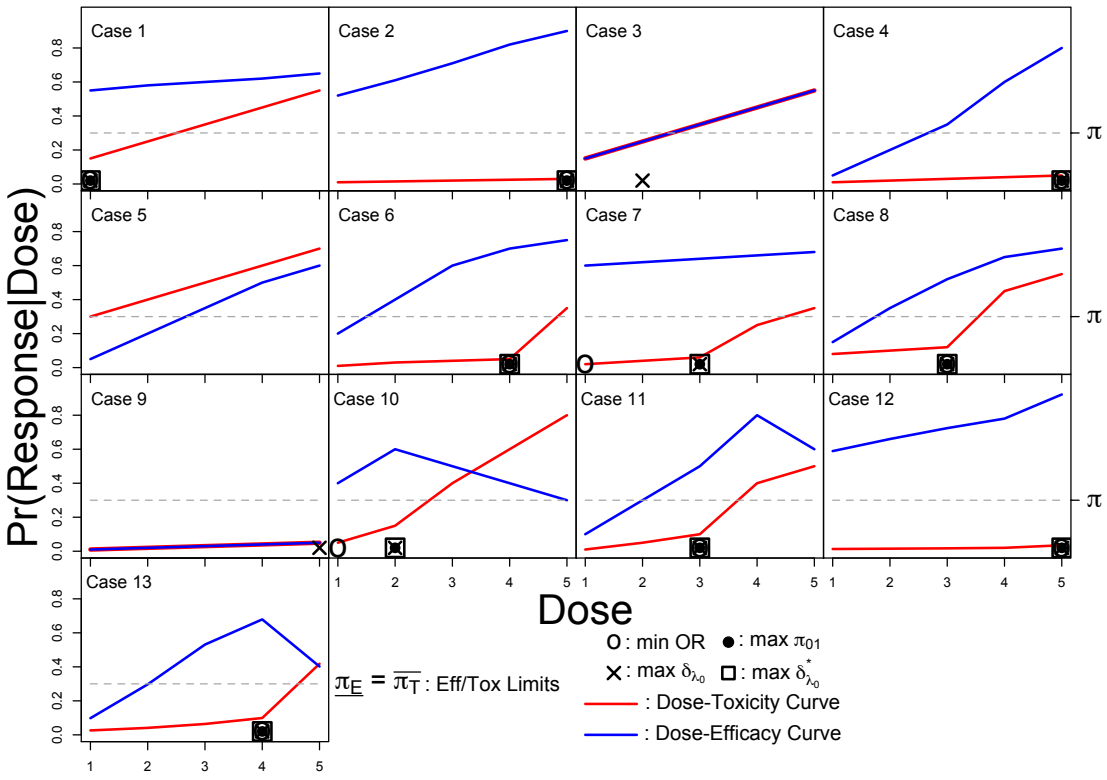

Figure 4.1: Thirteen combinations of dose-toxicity and dose-efficacy curves from Yin et al. (2006). The optimal dose based on different decision criteria are displayed above the pre-specified dose levels on the x-axis, and denoted as: (open circle) toxicity-efficacy odds ratio, (black dot) joint posterior probability of no toxicity with efficacy, (x) Zhang et al. (2006) decision rule with $\lambda = 0$, and (square) alteration: posterior probability of efficacy conditional on no toxicity, Zhang et al. (2006) decision rule with $\lambda = 0$. The dose-toxicity and dose-efficacy curves are represented with red and blue lines, respectively. The grey line displays the upper toxicity and lower efficacy limits for our posterior probabilities. **Scenario 1**: All five populations assume dose-response curves from Case 1. **Scenario 2**: Cases 3, 13, 13, 13, and 13 for populations 1, 2, 3, 4, and 5, respectively. **Scenario 3**: Cases 3, 9, 1, 8, 11. **Scenario 4**: Cases 6, 7, 8, 10, 13.

relatively little improvement (displayed as "OR HM"). A possible concern related to our design is that our more aggressive dose-finding algorithms might increase DLTs, but our results suggest that the rate of DLTs is similar across models both with and without HM. In addition to increasing the probability of correctly identifying the optimal dose, hierarchical modeling also increased the average number of patients treated at the optimal dose, although in this case, the hierarchical trinomial model (displayed as "Tri HM") displayed better performance than the PLR HM. These results highlight the benefits of hierarchical modeling: when the populations are homogenous and it is appropriate to share across populations, HM increases the probability of correctly identifying the optimal dose and the number of patients treated at the optimal dose with limited impact on the number of DLTs observed in the trial.

In Scenario 2, the last four populations are homogeneous (Case 13, where dose 4 is the optimal dose), while the first population has no dose level with an acceptable efficacy/toxicity trade-off (Case 3). This is a challenging scenario because the last four populations will encourage borrowing across populations but this could result in incorrectly borrowing strength from the first population, where no dose is acceptable. For the last four populations, HM greatly out-performs the independent models in correctly identifying the optimal dose and in treating more patients on average at the optimal dose, with limited impact on the probability of DLTs. This behavior is to be expected because the larger pooled sample size results in a more precise estimate of the BOD and allows individual populations to escalate more quickly than the independent designs. Comparing across hierarchical models, we see that the OR HM model has the best performance and largest improvement over its corresponding independence design, while the PLR HM and Tri HM have similar performance. For the first population, where no dose has an acceptable efficacy/toxicity trade-off, we see that HM decreases the probability of drawing the correct conclusion, with the two bivariate models exhibiting worse performance than the Tri HM model. This is to be expected due to the more stringent acceptability criterion used by the trinomial model. Finally, we note that while the HM designs decreased the probability of correctly concluding that no dose level is acceptable, HM had little impact on the number of DLTs observed with little increase over

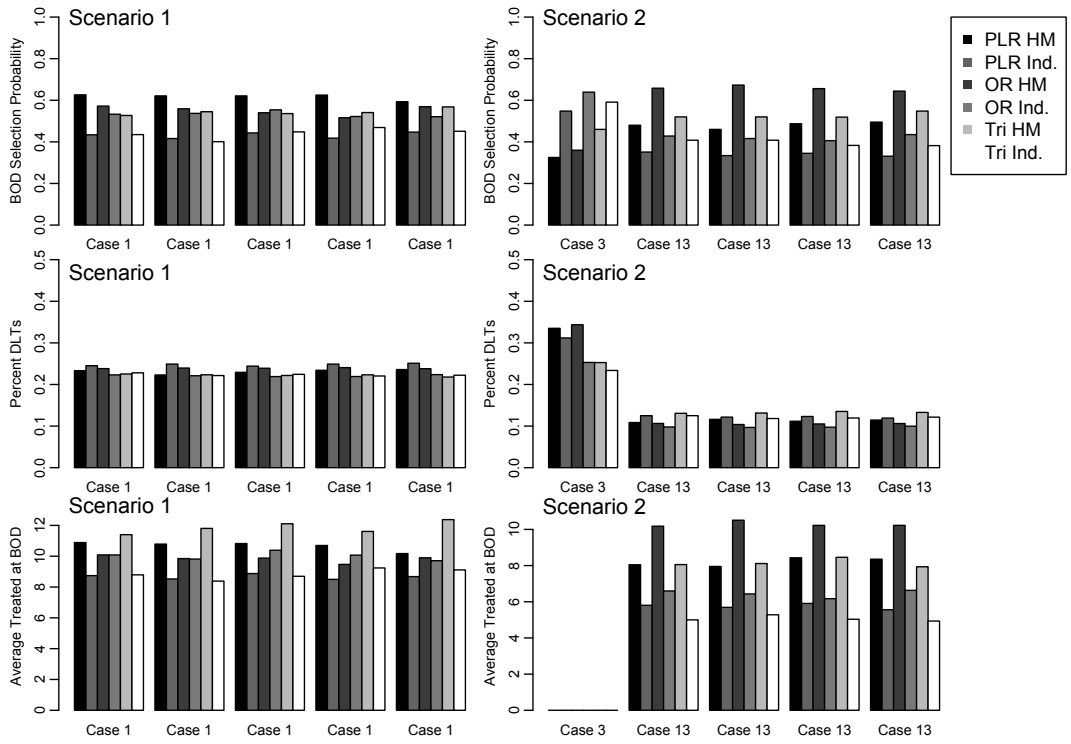

Figure 4.2: (Left column) Trial operating characteristics from 1000 simulated trials for Scenario 1 by population: (top) selection probability for the population-specific biologically optimal dose (BOD); (middle) percentage of dose-limited toxicities; (bottom) average number of patients treated at the population specific BOD, for the three hierarchical and independence designs. The first two bars (black and dark grey; labelled "PLR HM" and "PLR Ind.", respectively) present results using the parametric bivariate binary models. The next two bars (darker grey and grey; labelled "OR HM" and "OR Ind.", respectively) present results using the non-parametric bivariate binary models. The last two bars (light grey and white; labelled "Tri HM" and "Tri Ind.", respectively) present results for the parametric trinary model. For each scenario, the population's case-index used for data generation is presented on the x-axis. (Right column) Simulation results for Scenario 2.

the independence designs.

Results for Scenarios 3 and 4 can be found in Figure 4.3. Scenario 3 is another difficult case, where all doses have an unacceptable efficacy/toxicity trade-off in the first two populations (Cases 3 and 9, respectively), dose-level 1 is the optimal dose in the third population (Case 1) and dose-level 3 is the optimal dose for the last two populations (Cases 8 and 11, respectively). The results for Scenario 3 are consistent with our previous results. The three hierarchical models are more likely to correctly identify the optimal dose and treat more patients, on average, at the optimal dose when an optimal dose exists (last three populations) but the HM approaches are also more likely to incorrectly conclude that an optimal dose exists when no dose level has an acceptable efficacy/toxicity trade-off, although we again note that the impact on the total number of DLTs is minimal with the PLR HM having the highest DLT rate and Tri HM having the lowest DLT rate from among the three HM designs. Finally, comparing across hierarchical models, we see that the Tri HM design has the best performance of the three hierarchical models, treating more patients at the optimal dose and fewer patients above the optimal dose than the other two models.

In Scenario 4, there is modest heterogeneity in the optimal dose with the optimal dose varying from dose level 2 to dose level 4. We believe that this scenario best represents what we might expect to see in practice, for two reasons. First, investigators are advised to select their dose range such that the optimal dose is likely to be an intermediate dose based on pre-clinical data. Second, this scenario represents the case where the optimal dose is similar across populations but there is some variability due to different background treatments for each population. In this case, we see that the hierarchical extensions are more likely to correctly identify the optimal dose and treat more patients, on average, at the optimal dose in all populations. Among the hierarchical designs, the trinomial model performs the best across all populations. Again, HM has only a modest impact on the DLT rate compared to the independence designs, with the Tri HM again having the lowest DLT rate from among the three hierarchical designs and the PLR HM design having the highest. A possible solution to decreasing the DLT rate for

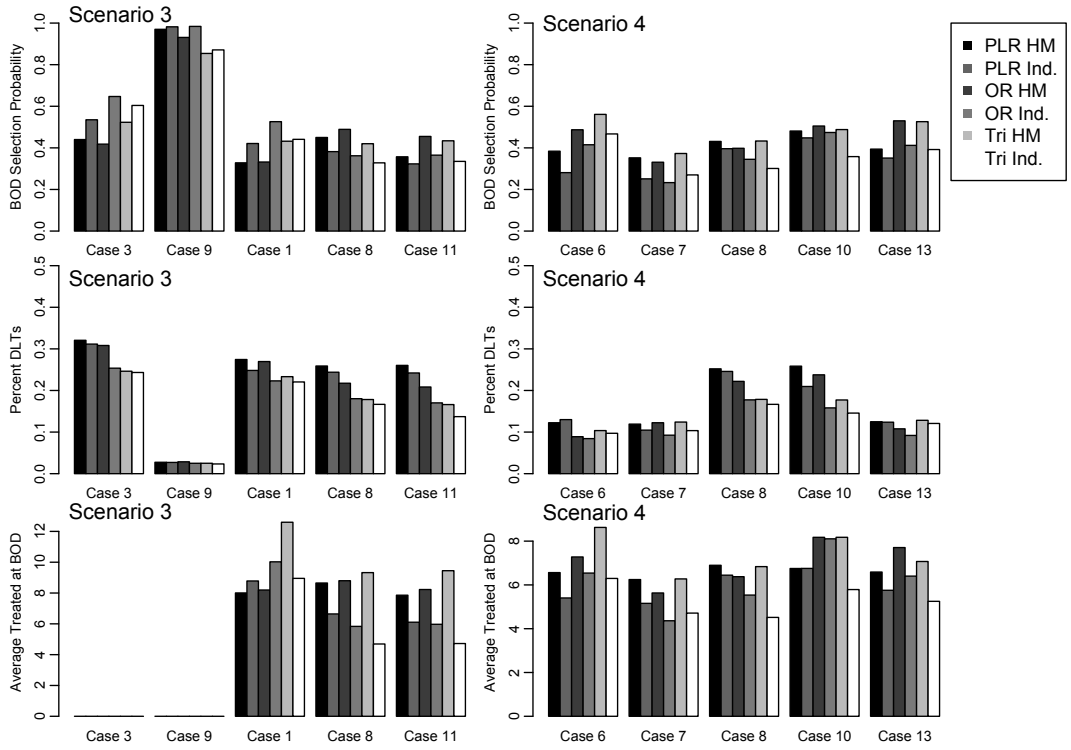

Figure 4.3: (Left column) Trial operating characteristics from 1000 simulated trials for Scenario 3 by population: (top) selection probability for the population-specific biologically optimal dose (BOD); (middle) percentage of dose-limited toxicities; (bottom) average number of patients treated at the population specific BOD, for the three hierarchical and independence designs. The first two bars (black and dark grey; labelled "PLR HM" and "PLR Ind.", respectively) present results using the parametric bivariate binary models. The next two bars (darker grey and grey; labelled "OR HM" and "OR Ind.", respectively) present results using the non-parametric bivariate binary models. The last two bars (light grey and white; labelled "Tri HM" and "Tri Ind.", respectively) present results for the parametric trinary model. For each scenario, the population's case-index used for data generation is presented on the x-axis. (Right column) Simulation results for Scenario 4.

the PLR HM design is to increase γ_T when using the parametric bivariate model but the impact of increasing γ_T on the other operating characteristics is a potential concern.

In summary, our simulation results highlight the benefits of HM in Phase I-II dose-finding trials. Incorporating HM resulted in an increased probability of correctly identifying the optimal dose and treating more patients at the optimal dose than the independent designs, with only a modest increase in the probability of DLT. Comparing the three hierarchical models discussed in Section 4.2, the trinomial hierarchical model resulted in the best trade-off between correctly sharing information across populations when the populations were homogeneous and over-sharing when the populations were heterogeneous. In addition, the trinomial model treated more patients at the optimal dose than the other two models, in most cases. In contrast, our simulation results suggest that implementing HM in the bivariate binary models and the parametric model, in particular, is not ideal. Both models were less likely to correctly declare all doses futile or unsafe and resulted in more DLTs when populations were heterogeneous compared to the trinomial model. Furthermore, they did not provide substantially more benefits than the trinomial model when populations were homogeneous.

4.5 Discussion

In this chapter we discussed HM for sharing information across populations in Phase I-II clinical trials. First, we presented two bivariate models, one parametric and one non-parametric, for modeling the dose-response relationship for efficacy and toxicity. Dose-finding using these models implemented the acceptability criteria defined by Thall and Cook (2004) to identify acceptable dose-levels with the optimal dose defined as the one maximizing the probability of efficacy with no toxicity (Yin et al., 2006). Next, we presented a hierarchical extension of the trinary outcome model proposed by Zhang et al. (2006), which combined the two binary outcomes into a single, trinary outcome. Reducing the two binary outcomes to a single trinary outcome precluded the direct application of acceptability criteria defined by Thall and Cook (2004) and instead we adapted the decision functions proposed by Zhang et al. (2006) for identifying the optimal dose with the trinary model.

Our simulation results suggest that the two hierarchical bivariate outcome models outperformed the trinary model when the populations are homogeneous and, in particular, the non-parametric bivariate model performed very well when the populations are homogeneous and the optimal dose is one of the higher dose-levels. On the other hand, the two bivariate outcome models did not perform as well when the populations were heterogeneous, and performed particularly poorly when no dose was acceptable. In contrast the trinary model emerged as simpler and, as a result, exhibited more consistent performance than the other two models.

The methods presented in Section 4.2.1 implement the design parameters from Yin et al. (2006). Our results suggest that additional tuning is required for the parametric bivariate model, especially for the toxicity acceptability criterion, γ_T. As with any Bayesian analysis, the results presented in Section 4.4.2 depend on the prior distribution and prior dose-response skeleton. While the trinomial model's prior skeleton corresponds to a higher optimal dose level a priori, this method exhibits the best performance with respect to preventing escalation to overly toxic dose-levels. In the future, we would like to investigate more prior distributions for the pooling variance parameter of the hierarchical trinomial model.

While our primary interest was in the performance of the three hierarchical models, the comparison of the three independent designs is also of interest because, to the best of our knowledge, these three models have never been compared using the same scenarios. While the dose-finding algorithms are different, it is interesting to note how the simpler yet conservative method performs against the more complex and flexible methods. Comparing both dose-finding designs under the same scenarios motivated the alteration to Zhang et al. (2006)'s decision function defining acceptable doses. We incorporate an additional criterion for the probability of efficacy conditional on no toxicity. It would be interesting to further investigate altering the minimum response rate for efficacy conditional on no toxicity and the acceptability threshold, i.e., a smaller quantile of the posterior for the probability of efficacy conditional on no toxicity.

Chapter 5

Conclusion

5.1 Summary of Major Findings

In this thesis, we developed Bayesian methods for adaptive Phase I and Phase I-II clinical trials. First, in Chapter 2 we considered a Phase I-II clinical trial of therapeutic cancer vaccines. We proposed a two-stage, randomized design for identifying the optimal vaccination schedule from a set of multiple candidate schedules. Stage one determines which schedules, if any, satisfy a minimum clinical performance level using the DLT and immune response rates. Schedules that achieve the minimum level of clinical performance in stage one were compared by the expected magnitudes of immune response to determine the optimal schedule. If an optimal schedule cannot be determined after stage one, Bayesian predictive probabilities were used to determine the number of subjects needed to achieve a conclusive result in stage two. Our simulation results illustrated that incorporating the additional endpoint for the magnitude of immune response improves our ability to identify the optimal schedule, and our two-stage approach using Bayesian predictive probabilities dramatically increased the probability of achieving a conclusive result.

In Chapters 3 and 4, we proposed Bayesian adaptive Phase I and Phase I-II trial designs that use hierarchical modeling to share information across potentially heterogeneous populations. We first considered Phase I trial designs and discussed hierarchical extensions of three dose-toxicity models that are commonly used in Phase I clinical trials. In

addition, we proposed three novel dose-finding guidelines that facilitate dose-escalation in the presence of hierarchical modeling while protecting safety. Our simulation results suggested that hierarchical modeling often achieves a more precise estimate of each population's MTD and more patients treated at each population's MTD, while maintaining patient safety when implemented with the proposed dose-finding guidelines. The extension of the under-parameterized power model provided the best trade-off between borrowing strength when populations are homogeneous and robustness when populations are more heterogeneous. Next, we proposed hierarchical modeling in the context of Phase I-II clinical trials using one of the previously proposed dose-finding guidelines. The simple model that combines efficacy and toxicity into a single endpoint appears to performed best when using hierarchical modeling. These results further support the use of under-parameterized hierarchical models in early phase oncology trials. The proposed dose-finding guidelines did not increase the rate of DLTs when applied to Phase I clinical trials, although our results suggest that hierarchical models are more aggressive than the independent models when applied in Phase I-II clinical trials. Compared to the independence design, hierarchical Phase I-II designs display a dramatic increase in the number of patients treated at each population's optimal dose but at the cost of treating slightly more above each population's optimal dose. Whether this behavior is due to the extra complexity as a result of considering both efficacy and toxicity or over-sharing as a result of hierarchical modeling should be investigated further. Regardless, our results are promising and support the use of hierarchical modeling to share information across populations in early phase dose-finding trials.

5.2 Future Work and Considerations

We used hierarchical modeling to borrow strength across similar sources of information in Phase I and Phase I-II clinical trials. This is especially important given the limited sample sizes in early phase clinical trials. Our simulation results illustrate that hierarchical modeling can dramatically improve estimation and increase the number of patients receiving an active dose, with little impact on patient safety. Combining this approach with the randomized, two-stage design proposed in Chapter 2 when more than

two vaccination schedules are being evaluated is a subject worthy of further investigation.

Hyperparameters for our prior distributions were set by using simulation to compare trial operating characteristics under various prior specifications and identifying the set of hyper-parameters that provided the most robust performance. This is a common approach when implementing Bayesian adaptive trials designs; Lee and Cheung (2011) and Zhang et al. (2013) propose more formal approaches to specifying hyperparameters in early phase Bayesian adaptive clinical trials. A potential future extension of our work would be to adapt their approach to the designs proposed in this dissertation, which could potentially improve robustness and precision. Finally, evaluating our priors in terms of the prior effective sample size is a topic of particular interest. Following the work of Morita et al. (2012), we can approximate the number of patients required to achieve our prior precision under our proposed dose-finding algorithm. Prior effective sample size is an intuitive approach for quantifying the informativeness of a prior distribution and translating our prior specification to a metric more easily quantifiable to clinicians. This may help encourage the use of our proposed adaptive designs.

References

Babb, J., Rogatko, A., and Zacks, S. (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine* **17,** 1103–1120.

Barker, A., Sigman, C., Kelloff, G., and et al. (2009). I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics* **86,** 97–100.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10,** 1281–1312.

Berry, S. M., Broglio, K. R., Groshen, S., and Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase II oncology clinical trials. *Clinical Trials* **10,** 720–734.

Berry, S. M., Carlin, B. P., Lee, J. J., and Muller, P. (2010). *Bayesian Adaptive Methods for Clinical Trials*, volume 38. CRC Press.

Braun, T. (2002). The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Controlled Clinical Trials* **23,** 240–256.

Braun, T. M. (2014). The current design of oncology phase I clinical trials: progressing from algorithms to statistical models. *Chinese Clinical Oncology* **3,**.

Braun, T. M. and Wang, S. (2010). A hierarchical Bayesian design for phase I trials of novel combinations of cancer therapeutic agents. *Biometrics* **66,** 805–812.

Braun, T. M., Yuan, Z., and Thall, P. F. (2005). Determining a maximum-tolerated schedule of a cytotoxic agent. *Biometrics* **61,** 335–343.

Cheung, Y. K. and Chappell, R. (2000). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **56,** 1177–1182.

Chow, S.-C., Chang, M., et al. (2008). Adaptive design methods in clinical trials - A review. *Orphanet Journal of Rare Diseases* **3,** 169–90.

Chow, S.-C., Chang, M., and Pong, A. (2005). Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics* **15,** 575–591.

Chow, S.-C. and Corey, R. (2011). Benefits, challenges and obstacles of adaptive clinical trial designs. *Orphanet Journal of Rare Diseases* **6,** 79.

Cunanan, K. and Koopmeiners, J. S. (2014). Evaluating the performance of copula models in phase I–II clinical trials under model misspecification. *BMC Medical Research Methodology* **14,** 51.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* pages 909–917.

Durbec, J. and Sarles, H. (1978). Multicenter survey of the etiology of pancreatic diseases. *Digestion* **18,** 337–350.

Faries, D. (1994). Practical modifications of the continual reassessment method for phase I cancer clinical trials. *Journal of Biopharmaceutical Statistics* **4,** 147–164.

Food and Drug Administration et al. (2010). Guidance for industry adaptive design clinical trials for drugs and biologics. `http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790`.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., and Pinheiro, J. (2006). Adaptive designs in clinical drug development: An executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics* **16,** 275–283.

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

Gasparini, M. and Eisele, J. (2000). A curve-free method for phase I clinical trials. *Biometrics* **56,** 609–615.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1,** 515–534.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88,** 881–889.

Ghosh, S. K., Mukhopadhyay, P., and Lu, J.-C. J. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference* **136,** 1360–1375.

Goodman, S. N., Zahurak, M. L., and Piantadosi, S. (1995). Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine* **14,** 1149–1161.

Gooley, T. A., Martin, P. J., Fisher, L. D., and Pettinger, M. (1994). Simulation as a design tool for phase I/II clinical trials: An example from bone marrow transplantation. *Controlled Clinical Trials* **15,** 450–462.

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. *Biometrics* **66,** 532–540.

Iasonos, A. and O'Quigley, J. (2012). Interplay of priors and skeletons in two-stage continual reassessment method. *Statistics in Medicine* **31,** 4321–4336.

Iasonos, A. and O'Quigley, J. (2013). Design considerations for dose-expansion cohorts in phase I trials. *Journal of Clinical Oncology* **31,** 4014–4021.

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,** 465–477.

Ivanova, A. (2003). A new dose-finding design for bivariate outcomes. *Biometrics* **59,** 1001–1007.

Ivanova, A., Montazer-Haghighi, A., Mohanty, S. G., and et al. (2003). Improved up-and-down designs for phase I trials. *Statistics in Medicine* **22,** 69–82.

Jaki, T., Clive, S., and Weir, C. J. (2013). Principles of dose finding studies in cancer: A comparison of trial designs. *Cancer Chemotherapy and Pharmacology* **71,** 1107–1114.

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J., Gupta, S., et al. (2011). The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discovery* **1,** 44–53.

Koopmeiners, J. S. and Modiano, J. (2014). A Bayesian adaptive phase I-II clinical trial for evaluating efficacy and toxicity with delayed outcomes. *Clinical Trials* **11,** 38–48.

Korn, E. L. and Freidlin, B. (2011). Outcome-adaptive randomization: Is it useful? *Journal of Clinical Oncology* **29,** 771–776.

Korn, E. L., Midthune, D., Chen, T. T., Rubinstein, L. V., Christian, M. C., and Simon, R. M. (1994). A comparison of two phase I trial designs. *Statistics in Medicine* **13,** 1799–1806.

Kummar, S., Ji, J., Morgan, R., Lenz, H.-J., Puhalla, S. L., Belani, C. P., Gandara, D. R., Allen, D., Kiesel, B., Beumer, J. H., et al. (2012). A phase I study of veliparib in combination with metronomic cyclophosphamide in adults with refractory solid tumors and lymphomas. *Clinical Cancer Research* **18,** 1726–1734.

Lee, J. J., Chen, N., and Yin, G. (2012). Worth adapting? revisiting the usefulness of outcome-adaptive randomization. *Clinical Cancer Research* **18,** 4498–4507.

Lee, J. J. and Liu, D. D. (2008). A predictive probability design for phase II cancer clinical trials. *Clinical Trials* **5,** 93–106.

Lee, S. M. and Cheung, Y. K. (2011). Calibration of prior variance in the Bayesian continual reassessment method. *Statistics in Medicine* **30,** 2081–2089.

Legedza, A. and Ibrahim, J. (2001). Heterogeneity in phase I clinical trials: Prior elicitation and computation using the Continual Reassessment Method. *Statistics in Medicine* **20,** 867–882.

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,** 1681–1694.

Mehta, M. P., Wang, D., Wang, F., Kleinberg, L., Brade, A., Robins, H. I., Turaka, A., Leahy, T., Medina, D., Xiong, H., et al. (2015). Veliparib in combination with whole brain radiation therapy in patients with brain metastases: results of a phase 1 study. *Journal of Neuro-Oncology* **122,** 409–417.

Morita, S., Thall, P. F., and Müller, P. (2012). Prior effective sample size in conditionally independent hierarchical models. *Bayesian Analysis (Online)* **7,** 3.

National Cancer Institute (2006). Cancer vaccine fact sheet. `http://www.cancer.gov/cancertopics/factsheet/Therapy/cancer-vaccines`. [Accessed 17-July-2013].

Nebiyou Bekele, B. and Shen, Y. (2005). A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. *Biometrics* **61,** 343–354.

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

O'Quigley, J., Hughes, M. D., and Fenton, T. (2001). Dose-finding designs for HIV studies. *Biometrics* pages 1018–1029.

O'Quigley, J. and Paoletti, X. (2003). Continual reassessment method for ordered groups. *Biometrics* **59,** 430–440.

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

O'Quigley, J. and Shen, L. Z. (1996). Continual reassessment method: A likelihood approach. *Biometrics* pages 673–684.

Owonikoko, T. K., Dahlberg, S. E., Khan, S. A., Gerber, D. E., Dowell, J., Moss, R. A., Belani, C. P., Hann, C. L., Aggarwal, C., and Ramalingam, S. S. (2015). A phase 1 safety study of veliparib combined with cisplatin and etoposide in extensive stage small cell lung cancer: A trial of the ecogacrin cancer research group (e2511). *Lung Cancer* **89,** 66 – 70.

Paoletti, X. and Kramar, A. (2009). A comparison of model choices for the Continual Reassessment Method in phase I cancer trials. *Statistics in Medicine* **28,** 3012–3028.

Patterson, S., Francis, S., Ireson, M., Webber, D., and Whitehead, J. (1999). A novel Bayesian decision procedure for early-phase dose-finding studies. *Journal of Biopharmaceutical Statistics* **9,** 583–597.

Plummer, M. (2011). *rjags: Bayesian graphical models using MCMC*. R package version 3-10.

Reiss, K. A., Herman, J. M., Zahurak, M., Brade, A., Dawson, L. A., Scardina, A., Joffe, C., Petito, E., Hacker-Prietz, A., Kinders, R. J., et al. (2015). A phase I study of veliparib (abt-888) in combination with low-dose fractionated whole abdominal radiation therapy in patients with advanced solid malignancies and peritoneal carcinomatosis. *Clinical Cancer Research* **21,** 68–76.

Riviere, M.-K., Dubois, F., and Zohar, S. (2015). Competing designs for drug combination in phase I dose-finding clinical trials. *Statistics in Medicine* **34,** 1–12.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10,** 1–10.

Stangl, D. K. (1995). Prediction and decision making using Bayesian hierarchical models. *Statistics in Medicine* **14,** 2173–2190.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45,** 925–937.

Strauss, N. and Simon, R. (1995). Investigating a sequence of randomized phase II trials to discover promising treatments. *Statistics in Medicine* **14,** 1479–1489.

Stylianou, M. and Flournoy, N. (2002). Dose finding using the biased coin up-and-down design and isotonic regression. *Biometrics* **58,** 171–177.

Sutton, A. J., Abrams, K. R., Jones, D. R., Jones, D. R., Sheldon, T. A., and Song, F. (2000). *Methods for Meta-analysis in Medical Research.* New York: J. Wiley.

Sweeting, M., Mander, A., and Sabin, T. (2013). bcrm: Bayesian Continual Reassessment Method designs for phase I dose-finding trials. *Journal of Statistical Software* **54,** 13.

Thall, P., Fox, P., and Wathen, J. (2015). Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of Oncology* **26,** 1621–1628.

Thall, P. F. and Cook, J. D. (2004). Dose-finding based on efficacy-Toxicity trade-offs. *Biometrics* **60,** 684–693.

Thall, P. F., Nguyen, H. Q., Braun, T. M., and et al. (2013). Using joint utilities of the times to response and toxicity to adaptively optimize schedule–dose regimes. *Biometrics* **69,** 673–682.

Thall, P. F., Nguyen, H. Q., and Estey, E. H. (2008). Patient-specific dose finding based on bivariate outcomes and covariates. *Biometrics* **64,** 1126–1136.

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.

Thall, P. F., Simon, R., and Ellenberg, S. S. (1989). A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics* pages 537–547.

Thall, P. F. and Wathen, J. K. (2007). Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer* **43,** 859–866.

Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H., and Benjamin, R. S. (2003). Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine* **22,** 763–780.

Yao, T.-J., Begg, C. B., and Livingston, P. O. (1996). Optimal sample size for a series of pilot trials of new agents. *Biometrics* pages 992–1001.

Yao, T.-J. and Venkatraman, E. (1998). Optimal two-stage design for a series of pilot trials of new agents. *Biometrics* pages 1183–1189.

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,** 777–787.

Yin, G. and Yuan, Y. (2009). Bayesian model averaging Continual Reassessment Method in phase I clinical trials. *Journal of the American Statistical Association* **104,** 954–968.

Yuan, Y. and Yin, G. (2008). Sequential continual reassessment method for two-dimensional dose finding. *Statistics in Medicine* **27,** 5664–5678.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association* **64,** 131–146.

Zhang, J. and Braun, T. M. (2013). A phase I Bayesian adaptive design to simultaneously optimize dose and schedule assignments both between and within patients. *Journal of the American Statistical Association* **108,** 892–901.

Zhang, J., Braun, T. M., and Taylor, J. M. (2013). Adaptive prior variance calibration in the Bayesian continual reassessment method. *Statistics in Medicine* **32,** 2221–2234.

Zhang, W., Sargent, D. J., and Mandrekar, S. (2006). An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine* **25,** 2365–2383.

Zhong, W., Koopmeiners, J. S., and Carlin, B. P. (2012). A trivariate continual reassessment method for phase I/II trials of toxicity, efficacy, and surrogate efficacy. *Statistics in Medicine* **31,** 3885–3895.

Zohar, S. and Chevret, S. (2007). Recent developments in adaptive designs for phase I/II dose-finding studies. *Journal of Biopharmaceutical Statistics* **17,** 1071–1083.