ABSTRACT

Topics in Bayesian Adaptive Clinical Trial Design Using Dynamic Linear Models and Missing Data Imputation in Logistic Regression

Yuanyuan Guo, Ph.D.

Chairperson: Dean M. Young, Ph.D.

Conventional Phase II clinical trial designs usually employ a logistic regression model to analyze the efficacy of a new drug and, therefore, assumes a monotone dose-response relationship. Also, the logistic regression model requires the response to be categorical and, thus, is not applicable for continuous data. The traditional design in Phase II determines if a new drug will be further tested in Phase III based on only drug efficacy and allocates an equal number of patients to each dosage, ignoring dose efficacy. Because of the limitations of conventional clinical trial designs, new adaptive designs have been proposed by researchers to improve the flexibility and adaptability of conventional designs.

In Chapter Two we propose an adaptive Bayesian design that uses a bivariate normal dynamic linear model for a Phase II clinical trial, and we compare its performance to a Bayesian fixed or non-adaptive design. The proposed Bayesian adaptive design can be utilized for continuous data and can model various dose-response relationships. We remark that for many dose-response relationships, our proposed adaptive Bayesian design can use fewer patients to obtain a correct decision concerning a drug's efficacy than the Bayesian fixed design.

Missing data arises in almost all research; that is, part of the data are missing for a subject. A data analyst must decide how to cope with the missing data from among the numerous imputation methods that can be used. However, one might not know which imputation method is best. The objective of this study is to evaluate the efficacy of five imputation methods.

In Chapter Four, we have compared the performance of complete-data-only, single-mean imputation, conditional-mean imputation, multiple imputation by chained equations and hotdeck imputation methods for prediction of a logistic regression model, for the missing-completely-at-random and missing-at-random mechanisms. These five imputation methods yield different results for small sample sizes, and the difference decreases with an increasing sample size. Surprisingly, a single-mean imputation method performs as well as the multiple imputation methods compared here.

Topics in Bayesian Adaptive Clinical Trial Design Using Dynamic Linear Models and Missing Data Imputation in Logistic Regression

by

Yuanyuan Guo, B.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

_____

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____

Dean M. Young, Ph.D., Chairperson

_____

James D. Stamey, Ph.D.

_____

Jack D. Tubbs, Ph.D.

_____

David A. Kahle, Ph.D.

_____

Van Pham, Ph.D.

Accepted by the Graduate School
December 2015

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I wish to say a sincere and grateful "Thank you" to my advisor, Dr. Dean Young, for his guidance and support over the past three years. This dissertation would not have been possible without you. Thank you for your flexibility in scheduling and encouragement from the beginning to the end.

I would also like to express my deepest gratitude to my parents, Jin Guo and Guimei Zhu. Your endless love gives me the strength to chase my dreams and encourages me when I get weary. I could never achieve any successes without you. You have always been there for me.

Finally, I want to thank my loving husband and best friend, Dong. Thank you for supporting me throughout this process and for always believing in me more than I believed in myself. To my precious son, Elson, you have given me so many things to be thankful for since the day you arrived. You make me smile even on my worst days.

DEDICATION

To my husband, Dong, and my son, Elson

CHAPTER ONE

Introduction

1.1 An Introduction to Clinical Trial Design

A clinical trial is the process of testing safety and efficacy when a company is developing a new drug and is usually very expensive and time-consuming to conduct. Clinical trials are generally divided into four phases (I to IV). However, a new phase, 0, has been established by Food and Drug Administration (FDA). The five phases are:

- Phase 0: Determine if a drug will behave in a human as the pre-clinical testing indicated.
- Phase I: Determine the toxicity and side effects, and find the safe dosage levels of the drug using a small cohort of patients.
- Phase II: Study the drug's efficacy and safety with a larger cohort of patients.
- Phase III: Confirm the drug's efficacy and side effects with a large cohort of patients.
- Phase IV: Understand the drug's mechanism of action, such as the interaction with other drugs and long-term side effects.

Here, we are interested in Phase II of a clinical trial. Phase II studies introduce the new drug into patients having the disease for which the drug is being developed. Phase II contains the initial efficacy study in a clinical trial with the primary purpose of determining the dosage needed to successfully treat patients. Historically, the determination of the dosage in Phase II relies on only the efficacy. One usually

assumes that the dose-response relationship is monotonically increasing such that a higher dosage causes a higher efficacy response. However, this monotone, non-decreasing assumption for efficacy might not hold. For example, at first the response may increase when the dosage is increased, but after reaching a certain point, the response could begin to decrease and become toxic.

The dose-response model for a traditional Phase II clinical trial study is usually the logistic function, which describes the relationship between treatment effects and dosage level. The logistic model requires categorical responses while, in some cases, we have continuous responses to the new drug. For example, in some situations the treatment effect is not measured as failure (0) or success (1) but as a continuous variable. In the following chapter, we expand the binary dose-response curve to a possible non-monotone continuous curve that is more flexible. Also, we incorporate both efficacy and toxicity in the model. This modeling approach makes a Phase II study more similar to a combination of traditional Phase I and Phase II studies. Thus, our approach allows for a non-monotone dose-response relationship.

In traditional clinical trial design, an equal number of subjects or patients is assigned to each treatment arm to statistically determine the best dose. Moreover, subjects are assigned to each treatment arm throughout the length of the study, even if they are assigned to an arm with a toxic dosage. However, an adaptive design allows sample-size modification as data accumulates in the study. That is, an adaptive design provides flexibility and efficiency with a smaller total sample size and increases the probability of correctly answering the clinical question of interest. Thus, more researchers have an increased interest in adaptive designs (Kairalla, Coffey, Thomann and Muller, 2012). The Food and Drug Administration (FDA, 2010) has shared the following motivations of drug developers for using adaptive design rather than traditional trial design:

- Adaptive design provides the same information with fewer subjects or patients and, hence, is more efficient.

- Adaptive design increases the likelihood of success of the study objective.

- Adaptive design provides better estimates of the dose-response relationship, which, in turn, improves one's understanding of the treatment's effect.

The Bayesian dynamic linear model that we propose in Chapter Two allows for a feasible adaptive design. For the Bayesian approach, the posterior distributions for each parameter are updated for each treatment. Also, based on the information collected during the clinical trial, we can readily decide if the trial should continue or be terminated. In an adaptive design, additional patients are assigned to the treatments proportional to the probability of the treatment's efficacy, and the trial can be stopped once sufficient evidence for ceasing that trial is attained. In this dissertation, we present our proposed adaptive Bayesian design simultaneously incorporating the detection of both efficacy and toxicity as an alternate to the conventional fixed-patient trial design. Thus, our Bayesian adaptive design can save substantive money, time, and resources by assigning the patients to a more effective treatment and can also stop the trial early once conclusive evidence has been obtained. We conduct a Monte Carlo simulation in Chapter Three to study the performance of our proposed Phase II Bayesian adaptive approach and compare it to a traditional Bayesian fixed-patient allocation method.

1.2 An Introduction to Markov Chain Monte Carlo Simulation

The Metropolis Monte Carlo algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller and Teller, 1953) has been widely used by chemists and physicists after computers became popular and made extensive computation feasible. Simulation became a major tool among statisticians after about 1990. The Metropolis algorithm was generalized by Hastings (1970) and is now called the Metropolis-Hastings algorithm. The

Gibbs sampler is a special case of the Metropolis-Hastings algorithm introduced by Geman and Geman (1984). Currently, the majority of Bayesian MCMC computing is accomplished through iterative Monte Carlo methods, including the Metropolis-Hastings algorithm and Gibbs sampler.

The Metropolis-Hastings algorithm can become very difficult to apply for high dimensional problems. Thus, the Gibbs sampler, a simplified case of the Metropolis-Hastings algorithm, is the method we utilize to simplify our simulation slightly. The Gibbs sampler is more practical when the complete set of full conditional distributions for all unknown parameters can be achieved. Observations can be generated from each full conditional distribution, and those observations determine the joint distribution of all parameters, assuming only mild conditions. Tanner (1993) states that "if the joint density is positive over its entire domain then it is uniquely determined by the full conditionals." In Chapter three, we employ a Gibbs sampler in the simulation, given that the full conditional distribution for each parameter is achievable. The Bayesian framework allows us to incorporate information as additional patients are involved in the trial and to make decisions in real time during the trial by updating the posterior parameter distributions.

1.3 An Introduction to the Missing Data Problem for Logistic Regression

Missing data is a common problem in statistical data analysis that can significantly affect a data-analysis result. We generally consider four missing-data mechanisms. From the simplest to the most general (Gelman and Hill, 2006), they are as follows.

Missing-Completely-at-Random: For this mechanism, the probability of missingness is the same for all the observations. For example, each patient decides whether or not to tell his/her age by rolling a die and refuses to tell if a "3" shows up.

Missing-at-Random: Missing-at-random is a more general missingness mechanism in which the probability that a variable is missing depends only on available information.

Thus, a variable is missing-at-random if the probability of missingness of this variable depends only on the other fully recorded variables.

Missingness That Depends on Unobserved Predictors: This type of missingness depends on the values that have not been recorded so that the missingness is no longer "at random." Gelman and Hill (2006) have given an illustrated medical example that a patient is more likely to drop out of the study if a particular treatment causes discomfort. This missingness can only be random if the "discomfort" is measured and observed for all patients.

Missingness That Depends on the Missing Value Itself: When the probability of missingness depends on the variable itself, the data analysis is difficult. For example, people with higher earnings are less likely to reveal their salaries.

A well-performing course of action for analyzing missing data should be sought. Numerous methods exist for dealing with this issue. In Chapter Four we study various imputation methods based on the missing-completely-at-random and missing-at-random mechanisms by Monte Carlo simulation. We calculate several performance measurements of each imputation method and compare the efficacy for all of the considered imputation methods.

CHAPTER TWO

Bayesian Adaptive Design Using a Dynamic Linear Model Incorporating Toxicity and Efficacy in Phase II Clinical Trials

2.1 Motivation

Statistical adaptive clinical designs have been developed during the past few decades to evaluate the safety and efficacy of a new drug to treat patients in clinical practice. The new adaptive designs, both frequentist and Bayesian, are motivated by limitations of the traditional four-phase clinical design, which considers only the efficacy in a single-arm Phase II trial in the form of categorized binary outcomes. Here, we propose an adaptive Bayesian clinical trial design for Phase II to assign patient cohorts to treatments adaptively based on both efficacy and toxicity, and to utilize continuous outcomes directly without losing information because of categorization.

The remainder of this chapter is organized as follows. In Section 2.2 we give the background of previous studies related to clinical trial design. In Section 2.3 we describe the dynamic linear model and a Bayesian framework of the proposed adaptive design.

2.2 Previous Studies

Historically, the statistical design of a Phase I clinical trial has attempted to determine the maximum tolerable dose (MTD) by relying only on toxicity and ignoring efficacy. After the MTD of the treatment has been determined in Phase I, a Phase II clinical trial studies the drug's efficacy, assuming that a dose with acceptable toxicity has been determined and that higher dosage causes a higher response (Korn, Midthun, Chen, Rubinstein, Christian and Simon, 1994; Goodman, Zahurak, and Piantadosi, 1995). However, in some situations the optimal treatment benefit might be achieved at a dosage less than the MTD. Thus, the term Phase I/II trials usually refer to clinical trials incorporating efficacy and toxicity simultaneously. Gooley, Martin,

Fisher and Pettinger (1994) have designed a Phase I/II study to evaluate the number of T cells needed in an allogeneic marrow graft in order to avoid rejection while not causing an unacceptable risk of graft-versus-host disease. This design considered two dose-response curves and showed that benefits exist when both efficacy and toxicity are incorporated in the study.

Similarly, a strategy for dose-finding in clinical trials using two dose-response curves and incorporating both efficacy and toxicity was designed by Thall and Russell (1998). They proposed a design for conducting single-arm clinical trials to determine the treatment that satisfies both efficacy and toxicity requirements and that stops the trial early once sufficient evidence of efficacy has been obtained. This design can be considered as a combination of traditional Phase I and Phase II trials because it involves both dose-finding and evaluation of efficacy and toxicity. This strategy was further explained in dose-finding studies with a donor lymphocyte infusion (DLI) trial for acute leukemia by Thall, Estey, and Sung (1999).

Based on Thall and Russell's (1998) Bayesian adaptive design for trinomial outcomes, Zhang, Sargent, and Mandrekar (2006) have described an approach incorporating both efficacy and toxicity for dose-finding for a treatment in a Phase I trial. Also, Zhang et al. (2006) have used a continuation-ratio model rather than a proportional-odds model and, hence, they have used different dose-selection criteria and stopping rules.

One question that must be answered before the Phase I/ II trial design is applied in drug development is, "What rules should one apply for dose-finding when both efficacy and toxicity are incorporated in the trial?" Traditional clinical trial design determines the optimal dosage based only on toxicity (Phase I) or efficacy (Phase II). Thall and Cook (2004) have presented a model-based Bayesian method for dose-finding in Phase I/II clinical trials based on efficacy-toxicity trade-offs. Their method accommodates trials with trinary outcomes, where efficacy and toxicity are

disjoint, and trials with bivariate binary outcomes where both events might occur. The method constructs a family of contours characterizing the efficacy and toxicity trade-off from the target probabilities of efficacy and toxicity provided by a physician, and then evaluates the desirability of each dosage by using the contours. Thall, Cook, and Estey (2007) have illustrated this outcome-adaptive Bayesian procedure with a trial of a biologic agent for acute myelogenous leukemia to demonstrate that the method works in practice. These researchers also have conducted a simulation study to assess the design's average behavior.

The interest in clinical trial designs that simultaneously involves both efficacy and toxicity has increased because they can save substantial resources during a study. Hoering, LeBlanc, and Crowley (2011) have proposed a two-step dose-finding trial with both efficacy and toxicity assessed to determine the optimal dosage for a targeted agent. The steps consisted of the following: 1) Traditional Phase I trial design to estimate the MTD by assessing toxicity only; and 2) A proposed design with dosages set at and below the MTD to evaluate each dosage level for both efficacy and toxicity. This two-step design using a logistic model was compared to a traditional trial design, and the authors concluded that the proposed trial design provided greater certainty for correctly determining the optimal dosage level.

Adaptive clinical trial designs also provide more flexibility and efficiency including a smaller total sample size, a more efficient treatment development process, and an increased chance of correctly answering the clinical question of interest (Kairalla et al., 2012). The FDA Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER) (2010) has released guidelines for adaptive design clinical trials in industry. This paper discussed the aspects of adaptive design that need special consideration and the information that should be included in the design for FDA review.

Leininger (2010) has proposed a Bayesian dynamic linear model to estimate the efficacy at each dosage level in Phase II of a clinical trial. His method incorporated an adaptive approach in the clinical trial design that allows for a non-monotone dose-response relationship, assigned the patients to more efficacious dosages, and stopped the trial early for success or futility. He also has compared an adaptive design to the traditional logistic model and found that the adaptive design significantly reduced the number of patients needed to detect effective drugs.

With the goal of jointly modeling efficacy and toxicity and finding the dosage based on the efficacy and toxicity trade-off, Koopmeiners (2013) has proposed a Phase I/II dose escalation study with delayed outcomes. He jointly modeled efficacy and toxicity as time-to-event outcomes, and applied the design to a Phase I/II clinical trial of a targeted toxin. The proposed adaptive design determined the optimal dosage at a rate similar to a study with binary efficacy and toxicity outcomes, but with substantial savings in cost and time.

Lewis, Viele, Broglio, Berry, and Jones (2013) have applied adaptive design in Phase II trials, which is similar to Leininger's (2010) design, to evaluate the addition of L-carnitine to the treatment of vasopressor-dependent septic shock. The design involved a dynamic dose-response model to improve the efficiency, allocated subjects adaptively, and stopped the trial early for success or futility. The authors demonstrated that the resulting trial determined the best dosage of L-carnitine efficiently and provided guidance concerning whether to continue development into Phase III.

2.3 A Dynamic Linear Model for Modeling the Dose-Response Curve

We next describe our proposed dynamic linear model (DLM) which provides flexibility in dose-response modeling that incorporates an adaptive design allowing adaptive subjects' allocation to treatments and early stopping rules for success or futility.

Generally DLMs are defined by two pdfs:

$$f(Y_t|\theta_t) \text{ and } g(\theta_t|\theta_{t-1}),$$

where Y_t is the observation at point t, θ_t is the vector of parameters at point t, and t is the dosage level. The evolution of $g(\theta_t|\theta_{t-1})$ allows smooth changes of the parameter in a DLM. In the dose-response case, $g(\theta_t|\theta_{t-1})$ indicates that the mean response to one dosage is related in some way to the means for its neighboring dosage levels.

Rather than assuming monotonicity for the dose-response relationship and forcing the response to be categorical, we allow the response to be continuous and expand the dose-response relationship to possible be non-monotone, thus incorporating more flexibility into our dose-response model. The dose-response curves presented in Figure 2.1 display types of dose-response relationships commonly encountered in clinical trials. The null response curve shows that the drug has no effect, regardless of the dosage. The slowly increasing curve illustrates an increasing effect as the dosage increases. The quickly increasing curve shows a dose-response relationship where the effect increases sharply when the dosage is increased at first, then holds constant at some maximum effect once the dosage is greater than a certain value. The non-monotone curve exhibits the non-monotonicity dose-response relationship where the drug effect increases as the dosage increases before a certain point and then decreases when the dosage is increased, thus indicating that the drug might become toxic to patients after a certain dosage point. We use these dose-response scenarios as possible dose-response relationships for efficacy in our study. For toxicity, we consider only the monotone relationship for simplification. That is, we assume that increasing dosage levels past a certain dosage level increases toxicity.

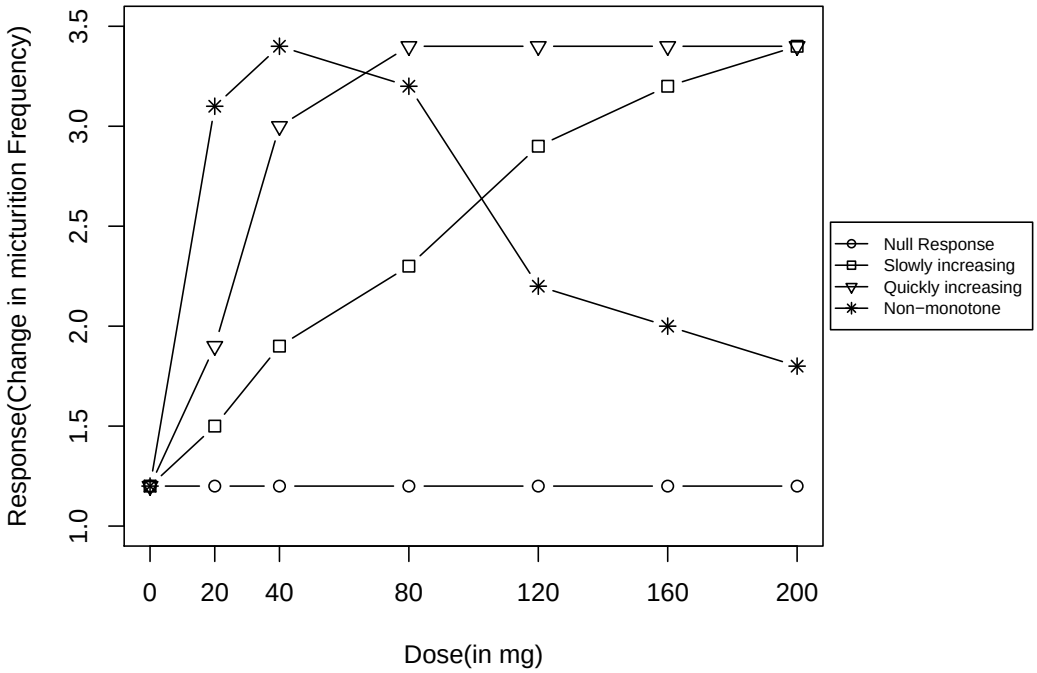

Figure 2.1: Possible dose-response scenarios for different drugs. The null response case is the control, while the slowly increasing, quickly increasing, and non-monotone curves represent effective drugs with varying dose-response relationships.

Generally, a dose-response model should relate the expected response at a given dosage to a set of parameters and covariates. An important class of models in the dose-response realm is the DLM, which is a simplified Gaussian process. The DLM does not restrict one to a monotone dose-response relationship like many other models, and it can easily handle a variety of possible dose-response relationships including non-monotone curves. Another advantage of the DLM is that one can easily implement it in a Bayesian framework.

2.3.1 The Likelihood Function

We use a bivariate normal DLM to incorporate both efficacy (X) and toxicity (Y). In the two-dimensional nonsingular case ($K = \text{rank}(\mathbf{\Sigma}) = 2$), the pdf of the vector $(x, y)'$ is

$$f(x_{ij}, y_{ij}) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_{ij}-\mu_{x_i})^2}{\sigma_x^2} + \frac{(y_{ij}-\mu_{y_i})^2}{\sigma_y^2} - \frac{2\rho(x_{ij}-\mu_{x_i})(y_{ij}-\mu_{y_i})}{\sigma_x\sigma_y}\right]\right),$$

where ρ is the correlation between X and Y, $\sigma_x > 0$ and $\sigma_y > 0$ and

$$\boldsymbol{\mu} \equiv \begin{pmatrix} \mu_{x_i} \\ \mu_{y_i} \end{pmatrix} \text{ and } \boldsymbol{\Sigma} \equiv \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}. \tag{2.1}$$

The likelihood function is

$$f(\boldsymbol{x}, \boldsymbol{y}) = (2\pi)^{-N} |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\Big\{ \frac{1}{2} \sum_{i=0}^{t} \sum_{j=1}^{n_i} [(x_{ij}, y_{ij})' - (\mu_{x_i}, \mu_{y_i})']' \times \boldsymbol{\Sigma}^{-1} [(x_{ij}, y_{ij})' - (\mu_{x_i}, \mu_{y_i})'] \Big\},$$

where $\boldsymbol{\Sigma}$ is given in (2.1). Also, we have

$x_{ij} \equiv$ toxicity of the j^{th} individual at the i^{th} dosage level;

$y_{ij} \equiv$ response of the j^{th} individual at the i^{th} dosage level;

$n_i \equiv$ number of patients who received the i^{th} treatment;

$t \equiv$ total number of treatments;

$N \equiv$ total number of patients tested;

$\mu_{x_i} \equiv$ the mean toxicity for the i^{th} treatment;

μ_{x_0} represents the mean toxicity of the placebo

$\mu_{y_i} \equiv$ the mean response for the i^{th} treatment;

μ_{y_0} represents the mean response of the placebo

$\sigma_x^2 \equiv$ the variance of individuals about the mean toxicity at each dosage level

and

$\sigma_y^2 \equiv$ the variance of individuals about the mean response at each dosage level.

2.3.2 Prior Distributions

To complete our Bayesian DLM, we next specify prior distributions on the parameters. Our Bayesian DLM assumes that the mean response at each dosage

has a normal distribution with its own mean but common variance. The conditional posterior distributions are updated as more information from additional patients that join in the trial becomes available.

Our prior set-up is similar to that of Leininger (2010) except that we use a bivariate model. The distribution of $(\mu_{x_i}, \mu_{y_i})'$ is a function of $(\mu_{x_{i-1}}, \mu_{y_{i-1}})$, and the prior distribution of (μ_{x_0}, μ_{y_0}) must be specified with initial parameter values because $(\mu_{x_{-1}}, \mu_{y_{-1}})$ does not exist. Also, d_i is included in the prior distribution so that the strength borrowed across dosages can be appropriately adjusted because the larger the value of d_i, the less the information borrowed. The support of these prior distributions matches that of the model parameters and allows us to determine closed-form conditional distributions for each unknown parameter.

We specify the prior distributions as

$$\begin{aligned}(\mu_{x_0}, \mu_{y_0})' &\sim N\left[(0,0)', \boldsymbol{B}\right], \\ (\mu_{x_i}, \mu_{y_i})' &\sim N\left[(\mu_{x_{i-1}}, \mu_{y_{i-1}})', d_i \boldsymbol{\Sigma}_\tau\right], \\ \boldsymbol{\Sigma} &\sim IW(\boldsymbol{R}, \nu),\end{aligned}$$

and

$$\boldsymbol{\Sigma_\tau} \sim IW(\boldsymbol{R}_\tau, \nu_\tau),$$

where $\nu,\ \nu_\tau > 0$, and

$s_i \equiv$ dosage for the i^{th} treatment;

$d_i \equiv \sqrt{s_i - s_{i-1}} =$ square root of the Euclidean distance between the current and previous dosage levels;

and

$$\nu = 2,$$

$$\nu_\tau = 2,$$

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\boldsymbol{R}_\tau = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and

$$\boldsymbol{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

to make the priors relatively non-informative. The notation IW represents the Inverse-Wishart distribution. We can see that the response at each dosage centers on the mean response of the previous dosage with no assumption of a monotonic dose-response relationship.

2.3.3 The Bayesian Adaptive Design

Our Bayesian adaptive design framework improves the flexibility and efficiency by adaptively assigning patients to dosages and by stopping the trial early for success or for futility. We describe the rationale of adaptive patient-treatment allocation and trial-stopping rules in the following subsections.

2.3.3.1 *An Adaptive Patient-Allocation Procedure.* An adaptive clinical trial design allows planned trial modifications based on data accumulated as additional patients are enrolled into the study. First, an initial group of patients is assigned to each dosage and to a control (placebo). Also, a preliminary dose-response curve is calculated and fit to the patients' measured responses. Next, all dosages are evaluated, and a new group of patients is assigned with a higher probability to those treatment arms with increased efficacy and non-increasing toxicity. Then the dose-response curve is recalculated using all of the available information. All dosages are reevaluated, and an additional group of patients is assigned to the dosages with a

higher probability of increased efficacy and decreased toxicity. In the trial, we continue to examine patients' responses and assign new groups of patients to dosages adaptively until a stopping rule is achieved.

Obviously, the dosages with large posterior-mean efficacy and small posterior-mean toxicity are typically assigned more patients. However, the variance of the posterior means of efficacy and toxicity also affects patient allocation because a large posterior variance indicates that we are highly uncertain about the location of the posterior mean. Thus, both the posterior means and posterior variances of efficacy and toxicity are considered in the allocation probability calculation for each dosage.

A fixed portion of each new cohort of patients is automatically allocated to the placebo arm so that the information concerning the placebo effect can be updated. For the non-placebo dosages, we calculate the probability of receiving additional patients as

$$P_i \equiv \frac{\sqrt{D_{Ei}^2 + D_{Ti}^2}}{\sum_{i=1}^{t} \sqrt{D_{Ei}^2 + D_{Ti}^2}}, \text{ for } i = 1, 2, \ldots, t,$$

where

$$D_{Ei} \equiv P(\mu_{Ei} > \text{ED95})\sigma_{\mu_{Ei}},$$

$$D_{Ti} \equiv P(\mu_{Ti} < \text{TD05})\sigma_{\mu_{Ti}},$$

and

$$\text{ED95} \equiv \mu_{E0} + 0.95(max\{\mu_{Ei}\} - \mu_{E0}),$$

$$\text{TD05} \equiv \mu_{T0} + 0.05(max\{\mu_{Ti}\} - \mu_{T0}).$$

We use $\sigma_{\mu_{Ti}} = 1$ because we do not want toxicity to dominate efficacy in a Phase II trial when allocating patients.

We then randomly allocate patients to each dosage using the corresponding allocation probabilities, and the conditional posterior distributions are updated after the responses of each new cohort of patients are observed. The allocation probability P_i is also recalculated to reflect the latest information, and then the clinical trial continues.

2.3.3.2 *Trial-Stopping Rules.* Patients are assigned to each dosage sequentially and adaptively depending on the dosage efficacy and toxicity. However, we also need stopping rules to decide when to stop the trial. These guidelines allow us to stop the trial early once sufficient information has been collected to draw a conclusion. For example, the collected data might strongly indicate that a dosage is not effective even though its toxicity is low. We might stop the trial at this point if we have enough evidence to conclude that this dosage is not effective. This stopping decision could save us substantive resources and cost for the clinical trial, and these subsequent patients can be assigned to a more effective treatment instead of an ineffective treatment.

To begin the trial, we first randomly assign a cohort of patients equally to each of the dosages arms and the placebo arm. We then use MCMC simulation to estimate the dose-response curve using the observed responses from the first cohort of patients. We then check our stopping rules described in Table 2.1.

Table 2.1: Adaptive design-stopping rules considered after each cohort of patients

Decision	Condition
Stop for success	Stop if $P(\mu_E^{ED95} \geq \mu_{E0}, \mu_T^{TD05} \leq \mu_{T0}) \geq U$ for any μ_{Ei} and μ_{Ti}
Stop for futility	Stop if $P(\mu_{Ei} \geq \mu_{E0}) \leq L$ for any μ_{Ei}
Stop for cap	$N \geq S$
Continue	None of the above conditions are met

The most likely average ED95 dosage, μ_E^{ED95}, is the dosage that has the highest posterior probability of having a mean efficacy above the ED95 level; The most likely average TD05 dosage, μ_T^{TD05}, is the one with the highest posterior probability of having a mean toxicity below the TD05 level. The ED95 and TD05 levels were defined in the last section. That is, $\mu_E^{ED95} \equiv \mu_{Ej}$ such that

$$P(\mu_{Ej} > \text{ED95}) = max\{P(\mu_{Ei} > \text{ED95})\} \text{ for } i = 1, 2, \ldots, t,$$

and $\mu_T^{TD05} \equiv \mu_{Tj}$ such that

$$P(\mu_{Tj} < \text{TD05}) = max\{P(\mu_{Ti} < \text{TD05})\} \text{ for } i = 1, 2, \ldots, t.$$

We stop a trial for success if $P(\mu_E^{ED95} \geq \mu_{E0}, \mu_T^{TD05} \leq \mu_{T0}) \geq U$ for some U where $0 < U < 1$.

The trial is stopped for futility if $P(\mu_{Ei} \geq \mu_{E0}) \leq L$ for some L where $0 < L < 1$. This criterion implies that no efficacious dosages were found, and we will not test those treatments. The toxicity is not of concern because the dosages with high toxicity have already been eliminated in Phase I, and the toxicity is irrelevant if the treatment is ineffective.

If at some point in the trial no dosages have either shown significant efficacy or been proved to be futile, we stop the trial because the maximum number of patients has been allocated. Stopping the trial for this reason is called *stopping for cap*. If none of the above conditions are met, the trial continues, and a new cohort of patients is allocated to the considered dosages as described in the previous section.

2.3.4 The Proposed Bayesian Adaptive Trial Using a Dynamic Linear Model

We start the trial with the first cohort of patients being assigned in equal number to the placebo and to six non-placebo dosage arms. The posterior distributions are updated using MCMC computation once the patient responses have been measured. Then, we check the stopping rules to determine if the trial will be stopped or will be continued. If none of the stopping rules are satisfied, the trial continues, and we allocate a new cohort of patients to the placebo and six dosage levels. The placebo automatically receives a fixed portion of the new cohort, and the remainder of the patients will be randomly allocated to one of the six dosage levels with probability proportional to the efficacy responses. Once the responses of the newest cohort of patients are processed, we then update the posterior distributions of the mean efficacy and mean toxicity again with all the available information. We next analyze the

stopping rules to decide if the trial will be continued. The trial continues in this pattern until a stopping rule has been satisfied. The proposed trial procedure can be summarized with the following algorithm:

Start of Trial: Allocate n_s patients to the placebo and non-placebo dosages.

(1) Measure the responses and use MCMC to update posterior distributions for unknown parameters $(\mu_{x_i}, \mu_{y_i})'$, $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_\tau$.

(2) Analyze the stopping rules.

- If none of the stopping rules are met, then
  - (a) Calculate p_i, the probability of the i^{th} dosage getting new patients for each dosage level.
  - (b) Randomly assign n_p patients from the new cohort to the placebo and the remaining patients to the non-placebo dosages with probability $p_i,\ i = 1, 2, \ldots, 6$.
  - (c) Repeat steps 1 and 2.
- If one of the stopping rules is met, proceed to the end of the trial.

End of Trial: Make a decision based on the stopping rules.

CHAPTER THREE

Monte Carlo Simulation

3.1 Simulation Design

We performed a simulation study using R and WinBUGS to evaluate the performance of a Bayesian adaptive DLM (BADLM) clinical trial design. A Monte Carlo simulation was conducted to compare the performance of BADLM clinical trial design with that of the Bayesian fixed patient-allocation DLM (BFPADLM) design. We also examined the dose-response curves estimated by our model to see if each curve is well estimated. The purpose of the chapter is to determine whether or not our DLM can capture the considered four types of dose-response curves and if the BADLM design performs better than the BFPADLM design.

We generated the data with normal random variables for each dose-response curve specified in Figure 2.1. We chose the four dose-response curves given in Figure 2.1 because they represent common dose-response relationships encountered in clinical trials. For toxicity, we considered only the monotone increasing dose-toxicity relationship. We examined two scenarios in terms of the correlation between efficacy and toxicity. The first scenario had a relatively small efficacy-toxicity correlation of $\rho = 0.2$, and the second scenario had a large efficacy-toxicity correlation of $\rho = 0.8$.

The prior distributions are specified as $\nu = 2,\ \nu_\tau = 2,$

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\boldsymbol{R}_\tau = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\boldsymbol{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

In each simulated clinical trial, 52 maximum batches of patients can be used, which lets us reach a cap of 600 patients. The clinical trials with a BADLM design and a BFPADLM design are conducted as follows:

a) a dynamic linear model with adaptive patient allocation

We began our simulated trial by assigning the first cohort of 84 patients to the placebo and to six non-placebo dosages with 12 patients assigned to each dosage. We then updated the posterior distributions of each dosage efficacy and toxicity parameter using the observed responses with MCMC computation. Next, we evaluated the stopping rules. If none of the stopping rules were satisfied, the trial continued with a new cohort of 10 patients allocated to the placebo and to the non-placebo dosages; 3 patients were automatically assigned to the placebo. For non-placebo dosages, we calculated P_i, the probability of receiving more patients of i^{th} dosage, and allocated patients to each non-placebo dosage with the corresponding allocation probability $P_i,\ i = 1, 2, \ldots, 6$. This process continued until a stopping rule was satisfied. The number of patients for each dosage and the standard deviation of the mean patients number for each dosage were also recorded.

b) a dynamic linear model with fixed patient-allocation

Rather than allocating patients to non-placebo dosages adaptively as in the BADLM clinical trial design, the simulation of BFPADLM assigned a fixed number of patients, 50, to each dosage. That is, all of the treatments including the placebo received only one cohort of patients in the trial with which to make a decision concerning whether or not the drug is effective.

We first determined the stopping rules via simulation under the scenario of small efficacy-toxicity and a BADLM design with 200 trials. Then, we simulated 1000 clinical trials using the determined stopping rules. The number of patients used at each dosage level and the standard deviation of the mean number of patients for each dosage level were reported and compared for the two competing designs. In

addition, we recorded the posterior mean response for each dosage and plotted the average 95% credible interval for each dose-response curve using 10000 iterations with a 1000-iteration burn-in period.

3.2 Conditional Distributions of Unknown Parameters

To estimate the DLM model proposed in section 2.3.1, we must solve for the conditional distributions of the unknown parameters. A detailed derivation of the conditional distributions of the unknown parameters are given in the Appendix. The conditional distributions are

$$(\mu_{x_0}, \mu_{y_0})' \sim N\Big\{[n_0\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}^{-1} + (d_1\boldsymbol{\Sigma}_\tau)^{-1}]^{-1}[n_0(\bar{X}_0)'\boldsymbol{\Sigma}^{-1} + \boldsymbol{\mu}_1'(d_1\boldsymbol{\Sigma}_\tau)^{-1}], \\ [n_0\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}^{-1} + (d_1\boldsymbol{\Sigma}_\tau)^{-1}]^{-1}\Big\},$$

$$(\mu_{x_t}, \mu_{y_t})' \sim N\Big\{[(d_t\boldsymbol{\Sigma}_\tau)^{-1} + n_t\boldsymbol{\Sigma}^{-1}]^{-1}[\boldsymbol{\Sigma}^{-1}n_t(\bar{X}_t) + (d_t\boldsymbol{\Sigma}_\tau)^{-1}\boldsymbol{\mu}_{t-1}], \\ [(d_t\boldsymbol{\Sigma}_\tau)^{-1} + n_t\boldsymbol{\Sigma}^{-1}]^{-1}\Big\},$$

$$(\mu_{x_i}, \mu_{y_i})' \sim N\Big\{[n_i\boldsymbol{\Sigma}^{-1} + (d_i\boldsymbol{\Sigma}_\tau)^{-1} + (d_{i+1}\boldsymbol{\Sigma}_\tau)^{-1}]^{-1}\left[\boldsymbol{\mu}_{i-1}'(d_i\boldsymbol{\Sigma}_\tau)^{-1} + n_i(\bar{X}_i)'\boldsymbol{\Sigma}^{-1}\right. \\ \left. + \boldsymbol{\mu}_{i+1}'(d_{i+1}\boldsymbol{\Sigma}_\tau)^{-1}\right], [n_i\boldsymbol{\Sigma}^{-1} + (d_i\boldsymbol{\Sigma}_\tau)^{-1} + (d_{i+1}\boldsymbol{\Sigma}_\tau)^{-1}]^{-1}\Big\},$$

$$\boldsymbol{\Sigma} \sim IW\Big\{N + \nu, \boldsymbol{R} + \sum_{i=0}^{t}\sum_{j=1}^{n_i}\left[(x_{ij}, y_{ij})' - (\mu_{x_i}, \mu_{y_i})'\right]\left[(x_{ij}, y_{ij})' - (\mu_{x_i}, \mu_{y_i})'\right]'\Big\},$$

and

$$\boldsymbol{\Sigma}_\tau \sim IW\Big\{\frac{t}{2} + \nu_\tau, \boldsymbol{R}_\tau + \sum_{i=1}^{t}\left[(\mu_{x_i}, \mu_{y_i})' - (\mu_{x_{i-1}}, \mu_{y_{i-1}})'\right]\left[(\mu_{x_i}, \mu_{y_i})' - (\mu_{x_{i-1}}, \mu_{y_{i-1}})'\right]'\Big\}.$$

Because all of the conditional distributions are in closed form, we can utilize a Gibbs sampling procedure to sample from the marginal posterior distribution of each parameter. Notice that the conditional distribution of $(\mu_{x_i}, \mu_{y_i})'$ is a function of $(\mu_{x_{i-1}}, \mu_{y_{i-1}})'$ and $(\mu_{x_{i+1}}, \mu_{y_{i+1}})'$, which implies that our Bayesian DLM allows one to borrow information from neighboring dosages. Also, $(\mu_{x_0}, \mu_{y_0})'$ and $(\mu_{x_t}, \mu_{y_t})'$ have

their own conditional distributions that are distinct from that of $(\mu_{x_i}, \mu_{y_i})'$ because $(\mu_{x_{0-1}}, \mu_{y_{0-1}})'$ and $(\mu_{x_{t+1}}, \mu_{y_{t+1}})'$ do not exist. This fact illustrates that $(\mu_{x_0}, \mu_{y_0})'$ and $(\mu_{x_t}, \mu_{y_t})'$ should have larger variances than $(\mu_{x_i}, \mu_{y_i})'$ because they borrow information from fewer neighboring dosages.

In our simulation study, we sampled from each conditional distribution in the order shown above. However, the ordering of the conditional distributions did not affect the simulation.

3.3 Determine Stopping Thresholds U and L

For stopping rules, we stop the trial early for success if

$$P(\mu_E^{ED95} \geq \mu_{E0}, \mu_T^{TD05} \leq \mu_{T0}) \geq U \text{ for any } \mu_{Ei} \text{ and } \mu_{Ti},$$

and we stop the trial early for futility if

$$P(\mu_{Ei} \geq \mu_{E0}) \leq L \text{ for any } \mu_{Ei} \text{ where } i = 1, 2, \ldots, t.$$

Under the scenario of a low efficacy-toxicity correlation of $\rho = 0.2$, we determine appropriate values for U and L by controlling Type I and Type II errors as in Leiningner (2010).

Assuming the tested drug was ineffective, i.e. none of the dosages were effective, we used an efficacy dose-response curve that was completely flat. We controlled the Type I error, which is the probability of concluding that the drug is effective when it is actually ineffective, to be no more than $\alpha = 0.05$ by choosing an appropriate value of L. The decisions of 200 trial simulations with different values of L are summarized in Table 3.1. We set $U = 0.0001$ for the determination of L and determined an appropriate value for U later.

Table 3.1: The effect of different values of L on Type I error for an ineffective efficacy dose-response curve. The Type I error rate of $\alpha \leq 0.05$ was achieved when $L = 0.999$.

L	Stopped for Success / Futility / Cap
0.999	0.020 /0.980 /0.000
0.99	0.085 /0.915 /0.000
0.95	0.230 /0.770 /0.000
0.90	0.360 /0.640 /0.000
0.80	0.605 /0.395 /0.000

The Type I error was less than 0.05 when $L = 0.999$ in Table 3.1, which satisfied the small Type I error requirement. Also, the trial length was reduced compared to the trial lengths of smaller values of L.

Next, we simulated to determine U by controlling the Type II error. Again we assumed an effective drug or a range of dosages and chose U to determine how often the drug dosages were declared ineffective when they were actually effective. We increased the values of U with $L = 0.999$ for slowly increasing efficacy dose-response curve. The results are summarized in Table 3.2.

Table 3.2: The effect of different values of U on Type II error for effective efficacy dose-response curves. The Type II error rate of $\beta = 0$ was achieved for all U's and attained the smallest cap when $U = 0.0001$.

U	Stopped for Success / Futility / Cap
0.2	0.075 / 0.000 / 0.925
0.1	0.165 / 0.000 / 0.835
0.01	0.555 / 0.000 / 0.445
0.001	0.820 / 0.000 / 0.180
0.0001	0.960 / 0.000 / 0.040

With $L = 0.999$, the Type II error rate of $\beta = 0$ was achieved for all considered values of U but with various caps. When $U = 0.2$, we stopped the trial for the cap as much as 0.925, which was undesirable. The caps decreased as U became smaller and reached 0.04 when $U = 0.0001$ without changing the Type II error rate. Therefore, we used $U = 0.0001$ and $L = 0.999$ throughout this paper because the desired Type I and Type II error rates were both achieved with this combination.

3.4 An Example Trial of Quickly Increasing Dose-Response Curve

To better understand the proposed BADLM clinical trial design, we first present an example of how a trial evolves for the quickly increasing dose-response curve where the drug becomes increasingly effective as the dosage increases in this section.

Figure 3.1 displays the estimated dose-response curve with 95% credible intervals of the trial in the first four stages. Each dose-response curve shows the estimated mean response of the patients at each dosage level with its 95% credible interval. Each dot above a dosage level represents that a patient has been randomly assigned to that dosage. For example, three dots are above 120 mg in Stage 1, thus indicating that three patients have been assigned to dosage 120 mg at that stage. The estimated dose-response curves in the first four stages show an increasing trend and that the credible intervals are getting more narrow as the stage number increases, especially for the dosage levels receiving the most patients. We also see that the dosage 200 mg was more effective than the competing dosages and, hence, that dosage received more patients in further stages of the trial. However, not enough evidence existed for us to make a decision through the first four stages.

Stages 5-8 in Figure 3.2 demonstrate the same quickly increasing dose-response curve as in the previous stages. The dosage of 200 mg received all of the patients in Stages 5 and 6, most of the patients in Stage 7, and all of the patients again in Stage 8. Obviously, the credible interval for a dosage of 200 mg is more narrow because it received more patients as the trial continued.

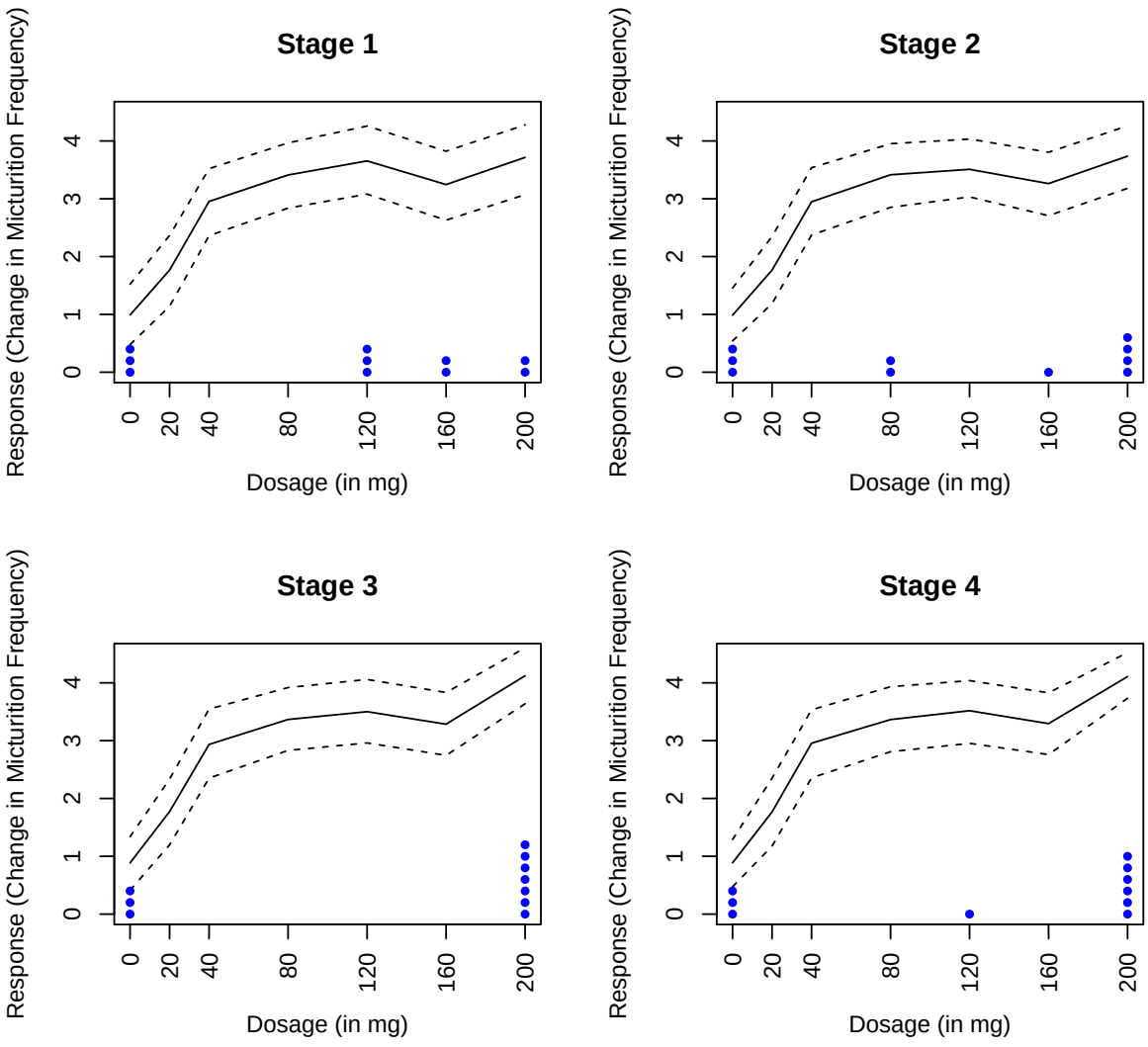

Figure 3.1: Stages 1-4 of a BADLM trial for a quickly increasing dose-response curve.

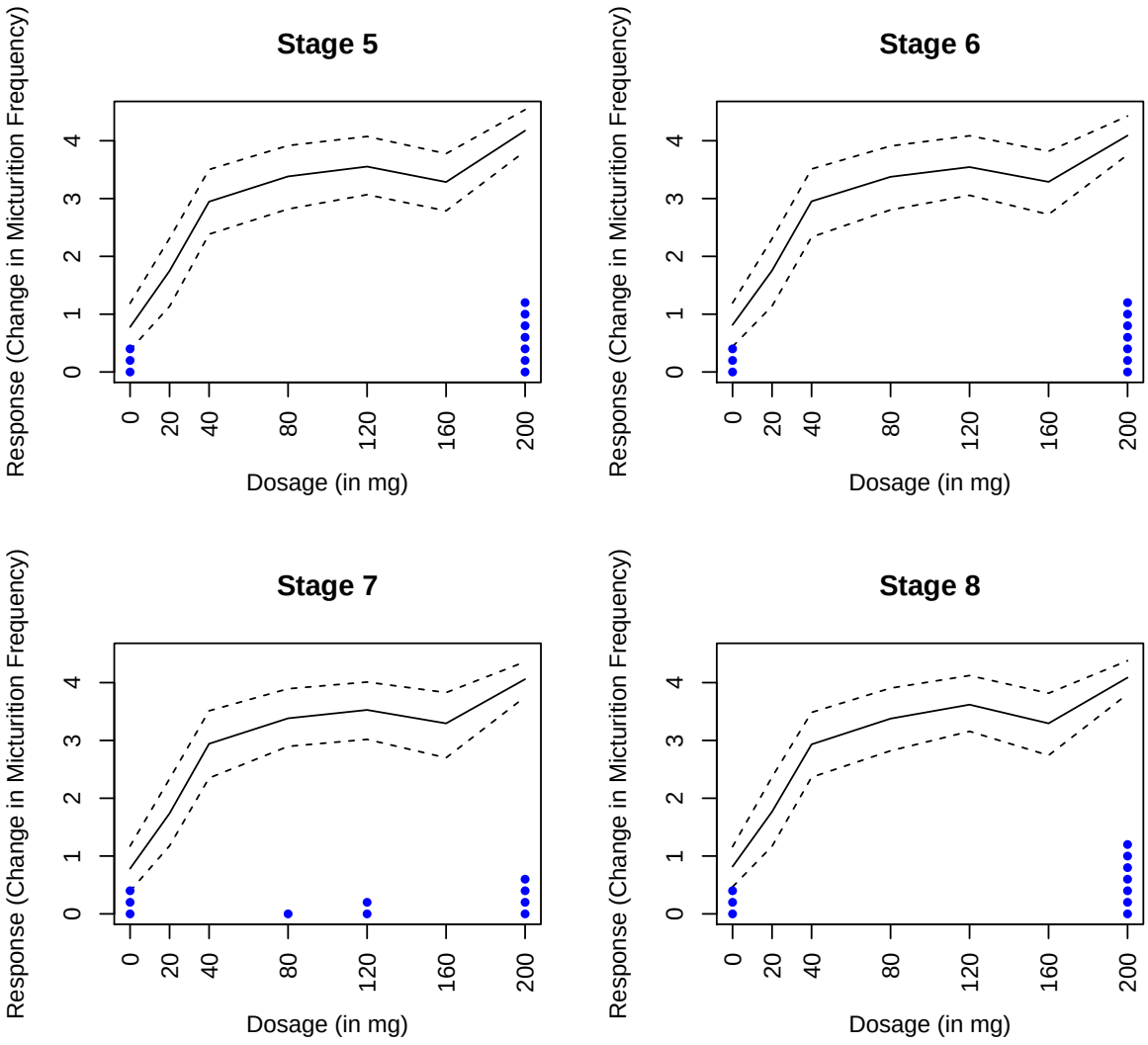

Figure 3.2: Stages 5-8 of a BADLM trial for a quickly increasing dose-response curve.

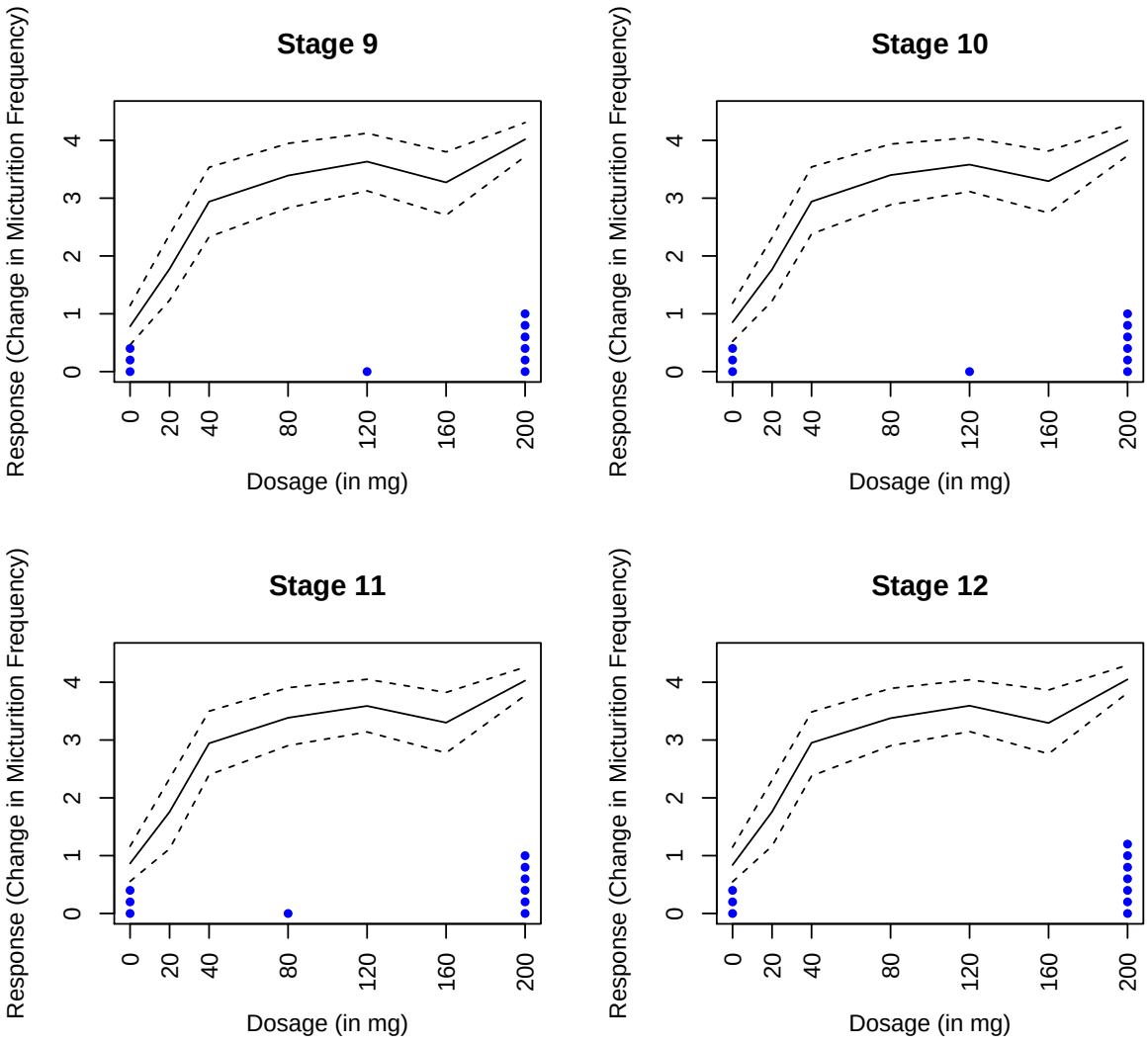

Figure 3.3: Stages 9-12 of a BADLM trial for a quickly increasing dose-response curve.

Figure 3.3 displays Stages 9-12 of the trial that are very similar to Stages 5-8 in Figure 3.2. The mean response at 200 mg was greater than the other dosages indicating that the dosage of 200 mg was more effective and, hence, the dose of 200 mg received more patients at each stage. The credible intervals continued to shrink as the trial advanced, especially for the dosage of 200 mg.

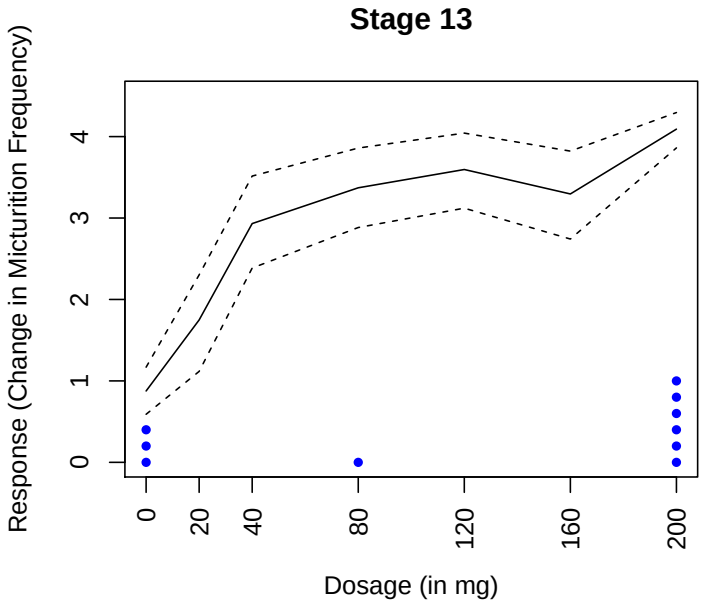

Figure 3.4: Stage 13 of a BADLM trial for a quickly increasing dose-response curve.

Stage 12, the last stage of the quickly increasing dose-response curve example trial, is shown in Figure 3.4. The 200 mg dosage level showed better efficacy than those of the other levels. The estimated dose-response curve appeared to be similar to the actual quickly increasing dose-response curve. This fact implied that the relationship between the dosage levels and response was correctly identified by our BADLM. Thus, the trial was stopped because sufficient evidence was collected to conclude that the drug was effective.

3.5 An Example Trial of Slowly Increasing Dose-Response Curve

We now examine a Phase II clinical trial having a slowly increasing dose-response curve for efficacy and a monotone increasing curve for toxicity. The example trial with three stages is shown in Figure 3.5. Again, each dose-response curve shows the estimated mean response of the patients at each dosage level with 95% credible intervals. Each dot above a dosage level represents that a patient was randomly assigned to that dosage.

From Figure 3.5 we see that our BADLM began to identify the slowly increasing trend in the dose-response at Stage 1 and our design allowed the dosage with a greater mean response to be assigned more patients. Notice that the dosage levels 160 mg and 200 mg received a large proportion of the patients as the trial continued because of their greater mean response. The credible intervals became more narrow as the trial continued, especially at the dosages with greater mean responses (160 mg and 200 mg). This trial took only four stages (the initial Stage and Stage 1-3) to conclude. We used 114 patients to stop the trial and make a correct decision that the drug was effective. Although the dosage level of 200 mg is the known ED95, our model detected that 160 mg and 200 mg are both effective and, hence, further analysis of the drug should be conducted for those two dosage levels.

3.6 An Example Trial of Non-monotone Dose-Response Curve

A non-monotone dose-response relationship describes an increasing drug effect as the dosage increases to a certain point, and then the dose-response curve decreases when the dosage is increased. Our BADLM was able to identify this type of dose-response relationship in early stages, as shown in Figure 3.6.

This example trial of Non-monotone dose-response curve took only four stages using 124 patients to stop the trial for success. As the dose-response curves in Figure 3.6 show, the response increased first as the dosage increased and then started to decrease after the dosage 40 mg. Hence, the dosage 40 mg received a larger proportion of patients at each stage. The adjacent dosages, 20 mg and 80 mg, were also assigned some patients to verify that they really were less efficacious than the dosage of 40 mg. Also, the credible intervals shrank as the trial evolved, especially at the dosage 40 mg.

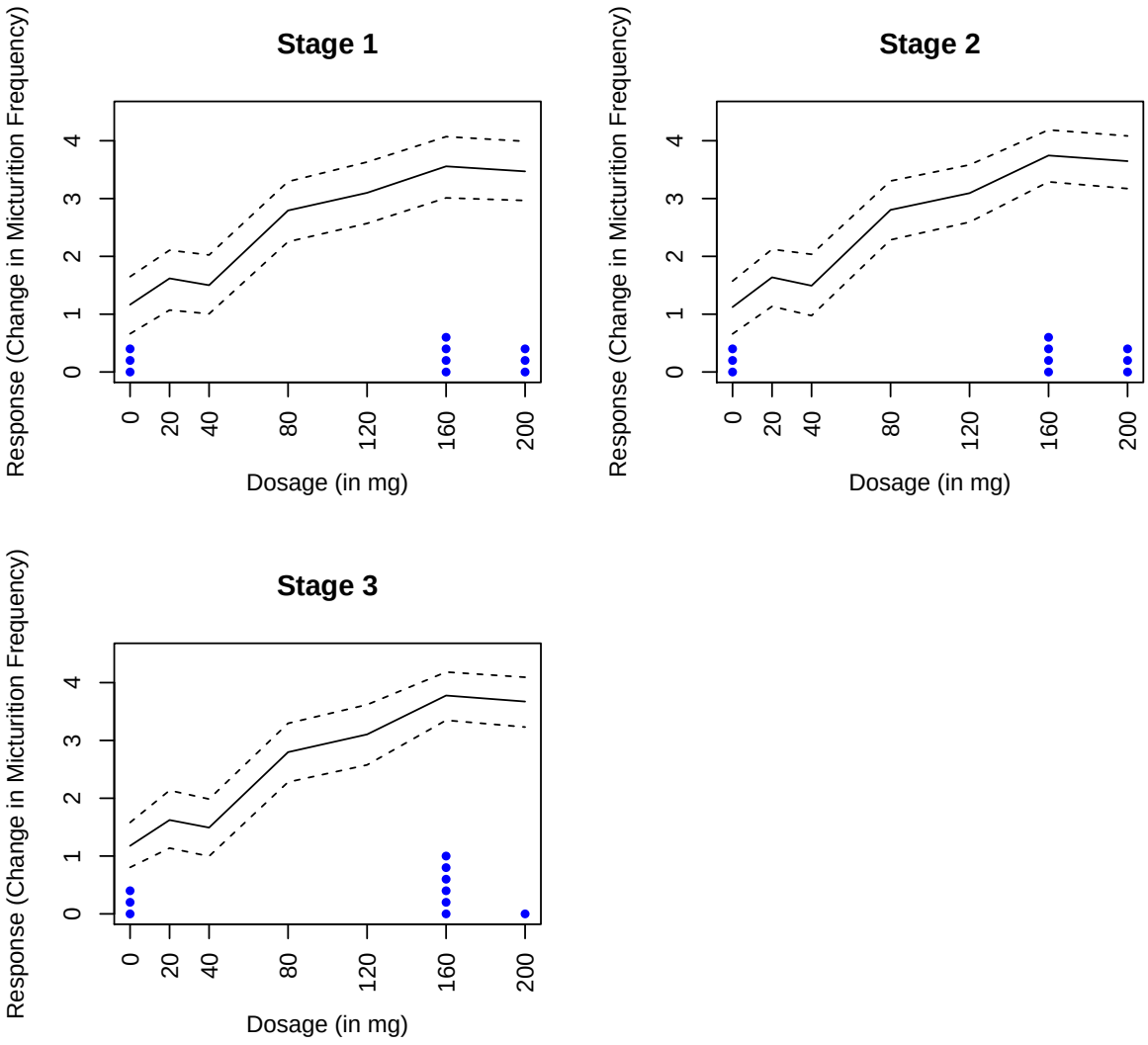

Figure 3.5: Stages 1-3 of a BADLM trial for a slowly increasing dose-response curve.

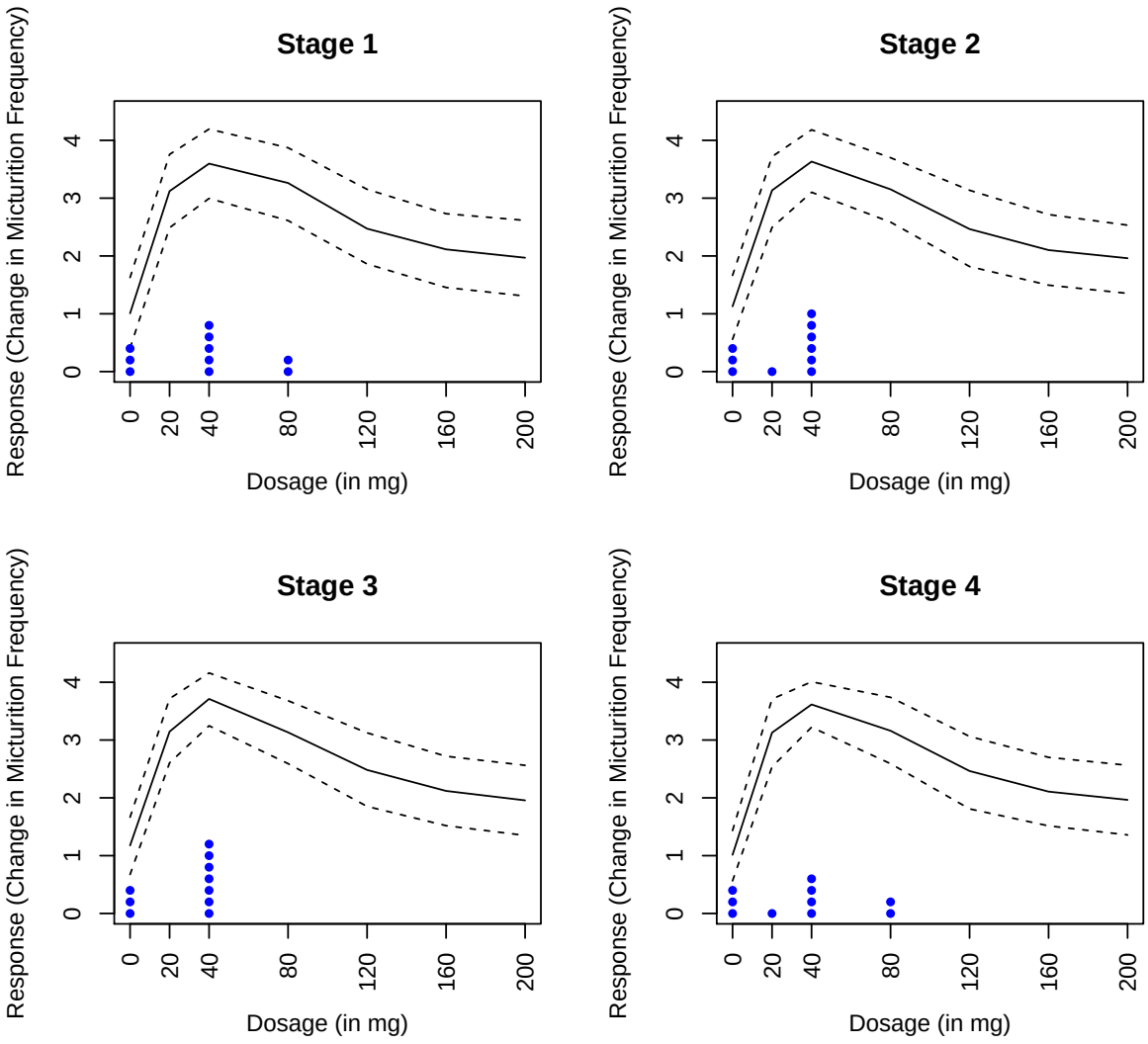

Figure 3.6: Stages 1-4 of a BADLM trial for the non-monotone dose-response curve.

3.7 Simulation Results

3.7.1 Stopping Decisions for a BADLM with Small Efficacy-Toxicity Correlation Scenario

Table 3.3 gives the stopping decisions for 1000 simulated trials for the BADLM for each dose-response curve under the scenario of small efficacy-toxicity correlation of $\rho = 0.2$. The values $U = 0.0001$ and $L = 0.999$ were fixed to attain a maximum of a 0.05 Type I error and a small Type II error. From Table 3.3, the Type I error from 1000 trials was $\alpha = 0.007$, which is less than our target value of $\alpha = 0.05$. For the three effective curves, the BADLM design was stopped for success for more than

0.94 of the simulated trials with a maximum Type II error of $\beta = 0.001$, which was the worst case for a non-monotone dose-response curve and caps under 0.06. So the chosen values of U and L gave us Type I and Type II errors that were less than the required upper bounds.

The stopping decisions for 1000 simulated trials for each of the four dose-response curves with a small efficacy-toxicity correlation for the BFPADLM design are summarized in Table 3.4. The BFPADLM design for the null dose-response curve made correct decisions for 0.99 of the simulated trials, which is very similar to the proportion of correct decisions for the BADLM design. However, for the three other types of dose-response curves considered here, the BFPADLM design was able to be stopped for correct decisions for approximately only half of the trials, and the proportion of trials stopped for cap increased correspondingly. The probability of a Type I error was $\alpha = 0.01$, and the probability of a Type II error was $\beta = 0$. Thus, in terms of the percentage of trials stopped for the correct decision, the BADLM design showed a very obvious advantage over the BFPADLM design.

Table 3.3: The BADLM-stopping decisions for 1000 simulated trials for four response curves under small efficacy-toxicity correlation scenario. The Type I error rate was $\alpha = 0.007$, and maximum Type II error rate was $\beta = 0.001$.

Dose-Response Curve	Stopped for Success / Futility / Cap
Null Response	0.007 / 0.993 / 0.000
Slowly Increasing	0.944 / 0.000 / 0.056
Quickly Increasing	0.947 / 0.000 / 0.053
Non-monotone	0.952 / 0.001 / 0.047

Table 3.4: The BFPADLM-stopping decisions for 1000 simulated trials for four response curves under small efficacy-toxicity correlation scenario. The Type I error rate was $\alpha = 0.01$, and Type II error rate was $\beta = 0$.

Dose-Response Curve	Stopped for Success / Futility / Cap
Null Response	0.010 / 0.990 / 0.000
Slowly Increasing	0.505 / 0.000 / 0.495
Quickly Increasing	0.509 / 0.000 / 0.491
Non-monotone	0.548 / 0.000 / 0.452

3.7.2 Patient Allocation for Small Efficacy-Toxicity Correlation Scenario

The number of patients used to make a decision for each trial is also an important metric in revealing the performance of a Phase II clinical trial. We have tracked the average number of patients at each dosage level, the total sample size per trial, and the standard deviation of each average sample size for our new BADLM design and the BFPADLM design summarized in Table 3.5.

Table 3.5 shows the average sample sizes per trial for each dose-response curve with both the BADLM and the BFPADLM designs when the correlation between efficacy and toxicity is small ($\rho = 0.2$). The BADLM design used a total number of 84 of patients to reach a correct decision of an ineffective treatment (null dose-response curve), while the BFPADLM design used a total number of 350 patients. Therefore, the BADLM clinical trial design could save a significant number of patients if we are experimenting with an ineffective drug treatment.

For an effective treatment, the adaptive design in Table 3.5 also showed a significant saving of patients by assigning more patients to the more effective treatments. For example, the non-monotone curve had an average of 12.009 patients assigned to dosage of 200 mg in the BADLM design, while a fixed number of 50 patients were used for the same dosage in the BFPADLM design. For the other dose-response curves,

the dosage of 20 mg for the slowly increasing dose-response curve received an average of 12.001 patients in the BADLM design versus 50 patients in the BFPADLM design. The sample-size difference in the two designs was also reflected in the average total number of patients used for each dose-response curve. The BADLM design used an average of 110.4 patients for the quickly increasing dose-response curve, while the BFPADLM design used a total of 350 patients to reach an even higher proportion of correct decision. We can also see the difference between the BFPADLM and BADLM designs in terms of the average total number of patients used for the quickly increasing curve as shown in Table 3.5.

In summary, considerably fewer patients were needed for each effective dose-response curve under the BADLM design compared to the BFPADLM design. The savings in terms of patients required were similar for the three effective (non-placebo) dose-response curves, and the sample size used for the ineffective (placebo) dose-response curve was even larger.

Table 3.5: Average sample sizes per trial for each curve for the BADLM and BFPADLM designs and small correlation between efficacy and toxicity ($\rho = 0.2$). (The known ED95 dosages are in blue).

Design	**Dose-Response Curve**	0 mg	20 mg	40 mg	80 mg
BADLM	Null Response	12(0)	12(0)	12(0)	12(0)
	Slowly Increasing	20.397(33.814)	12.001(0.032)	12.002(0.063)	12.065(0.887)
	Quickly Increasing	19.920(32.962)	12.000(0)	12.818(10.713)	**16.389(28.961)**
	Non-monotone	19.122(31.350)	15.801(26.715)	**19.296(39.896)**	17.497(34.141)
BFPADLM	All Curves	50(0)	50(0)	50(0)	50(0)

Design	**Dose-Response Curve**	120 mg	160 mg	200 mg	**Total**
BADLM	Null Response	12(0)	12(0)	12(0)	84(0)
	Slowly Increasing	14.111(19.975)	18.812(37.672)	**22.602(51.247)**	111.990(112.715)
	Quickly Increasing	**16.045(27.007)**	**17.060(31.606)**	**16.168(26.958)**	110.400(109.874)
	Non-monotone	12.008(0.134)	12.007(0.138)	12.009(0.285)	107.740(104.499)
BFPADLM	All Curves	50(0)	50(0)	50(0)	350(0)

3.7.3 Stopping Decisions of Dynamic Linear Model with Large Efficacy-Toxicity Correlation

We also conducted a similar Monte Carlo simulation except that we increased the efficacy-toxicity correlation from $\rho = 0.2$ to $\rho = 0.8$. Table 3.6 gives the stopping decisions for 1000 simulated trials for BADLM for each of the four dose-response curve under the scenario of a large efficacy-toxicity correlation of $\rho = 0.8$.

We can see from Table 3.6 that the Type I error from 1000 simulated trials was 0.005, which is somewhat less than the small efficacy-toxicity correlation scenario, and the Type II error was approximately zero. Hence, both Type I and Type II errors were sufficiently controlled. Compared to the small efficacy-toxicity correlation scenario, the percentage of correct decisions for a non-monotone curve slightly increased while the percentage of correct decisions of the slowly and quickly increasing dose-response curves decreased. The percentage of correct decisions for a slowly increasing dose-response curve dropped below 0.9 while stopping for cap increased correspondingly.

We also conducted a BFPADLM design simulation under large efficacy-toxicity correlation scenario ($\rho = 0.8$). Table 3.7 summarizes the results. The Type I and Type II error rates were $\alpha = 0.012$ and $\beta = 0$, respectively, which are less than the targeted error probabilities. However, the percentage of correct decisions for the slowly increasing dose-response curve dropped from 0.825 to 0.117, and the percentage of stopping for cap increased from 0.175 to 0.883 compared to the BADLM design. Similarly, for the quickly increasing dose-response curve, the percentage of correct decisions decreased to 0.446 compared to 0.914 for the BADLM design, and stopping for cap increased correspondingly. The percentage of correct decisions for the non-monotone dose-response curve also decreased to 0.524 with the proportion of cap of 0.476. This result is understandable because the stopping rules were determined with the small efficacy-toxicity correlation scenario and, therefore, the stopping rules might not be optimal when the efficacy-toxicity correlation is large.

Table 3.6: The BADLM-stopping decisions for 1000 trials for each response curve under large efficacy-toxicity correlation scenario. Type I error rate was $\alpha = 0.005$, and Type II error rate of $\beta = 0$ was achieved.

Dose-Response Curve	Stopped for Success / Futility / Cap
Null Response	0.005 / 0.995 / 0.000
Slowly Increasing	0.825 / 0.000 / 0.175
Quickly Increasing	0.914 / 0.000 / 0.086
Non-monotone	0.961 / 0.000 / 0.039

Table 3.7: The BFPADLM-stopping decisions for 1000 trials for each response curve under large efficacy-toxicity correlation scenario. A Type I error rate was $\alpha = 0.012$, and a Type II error rate of $\beta = 0$ was achieved.

Dose-Response Curve	Stopped for Success / Futility / Cap
Null Response	0.012 / 0.988 / 0.000
Slowly Increasing	0.117 / 0.000 / 0.883
Quickly Increasing	0.446 / 0.000 / 0.554
Non-monotone	0.524 / 0.000 / 0.476

3.7.4 Patient Allocation for Large Efficacy-Toxicity Correlation Scenario

Similar to the small efficacy-toxicity correlation scenario, a significant saving of patients was attained for the large efficacy-toxicity correlation scenario. Table 3.8 shows the average sample sizes per trial and the standard deviations of the mean number of patients for each type of dose-response curve for both the BADLM and BFPADLM designs when the correlation between efficacy and toxicity was large ($\rho = 0.8$).

The BADLM design and BFPADLM design showed large differences for all of the dose-response curves in terms of the number of patients used for each dosage. For the null dose-response curve, we found that we used an average of only 84 patients

with the BADLM design to reach a decision while the BFPADLM design used an average of 350 patients. For the non-monotone dose-response curve, the BADLM used an average of 103.470 patients, which is also considerably fewer than the 350 patients used for the BFPADLM design. For the quickly and slowly increasing dose-response curves, the BADLM also used fewer patients on average than the BFPADLM design. For the slowly increasing dose-response curve, the 200 mg dosage used the most patients with an average of 49.744, compared to all other dosages for which the average total number of patients used was 173.490 for the BADLM design. The BFPADLM design used an average of 350 patients for the slowly increasing dose-response curve, which is considerably more than the BADLM design required. Notice in Table 3.6 and Table 3.7 that even with many more patients used for a slowly increasing curve, the BFPADLM design obtained only 0.117 correct decisions out of 1000 trials, while the BADLM design obtained 0.825 correct decisions. Thus, the BFPADLM design performed worse than the BADLM design in terms of the percent of correct decisions made, although it used many more patients. Similarly, for the quickly increasing dose-response curve, the BFPADLM design used an average of 350 patients and obtained 0.446 correct decisions out of 1000 trials, while the BADLM used an average of 127.64 patients with 0.914 correct decisions. The difference in terms of the total number of patients used between two designs for a non-monotone curve was as significant as for the other dose-response curves. The percentage of correct decisions was 0.961 for the BADLM and 0.524 for the BFPADLM design.

As in the last section, the more effective dosages received more patients in the BADLM design. For example, the 200 mg treatment with the slowly increasing curve received an average of 49.74 patients, while the 20 mg treatment was assigned an average of only 12.14 patients under the BADLM design. The BFPADLM design assigned an equal number of patients to each treatment regardless of the efficacy of each treatment. For the small efficacy-toxicity correlation scenario in the last section,

the BADLM design used a smaller number of patients for each effective dose-response curve than the BFPADLM design and also reached more correct decisions result for a larger proportion of 1000 simulated trials.

Table 3.8: Average sample sizes per trial for each curve for the BADLM and the BFPADLM designs with large correlation between efficacy and toxicity ($\rho = 0.8$). (The known ED95 dosages are in blue).

Design	Dose-Response Curve	0 mg	20 mg	40 mg	80 mg
BADLM	Null Response	12(0)	12(0)	12(0)	12(0)
	Slowly Increasing	38.847(55.922)	12.144(1.995)	12.001(0.032)	12.202(1.890)
	Quickly Increasing	25.092(41.291)	10.002(0.045)	13.115(11.399)	**18.802(35.208)**
	Non-monotone	17.841(28.508)	18.107(37.326)	**16.414(28.734)**	15.097(322.245)
BFPADLM	All Curves	50(0)	50(0)	50(0)	50(0)

Design	Dose-Response Curve	120 mg	160 mg	200 mg	Total
BADLM	Null Response	12(0)	12(0)	12(0)	84(0)
	Slowly Increasing	16.470(22.323)	32.082(62.530)	**49.744(92.703)**	173.490(186.407)
	Quickly Increasing	**17.809(30.410)**	**19.690(37.644)**	**21.130(41.806)**	127.64(137.638)
	Non-monotone	12.010(0.190)	12.001(0.032)	12(0)	103.470(95.027)
BFPADLM	All Curves	50(0)	50(0)	50(0)	350(0)

3.7.5 Dose-Response Curve Estimation Comparison for Small Efficacy-Toxicity Correlation Scenario

We now compare the dose-response curves estimated under the BADLM and BFPADLM design. The four dose-response curves are plotted in Figure 3.7-3.10. Figure 3.7 shows two similarly shaped null dose-response curves with 95% credible intervals for the BADLM and BFPADLM designs. The smoothness of the two curves varies. We can see that for the dose-response curve under the BFPADLM design, the 95% credible intervals are narrower than those for the BADLM design. However, the BADLM design correctly identified futility with only 84 patients while the BFPADLM design used 350.

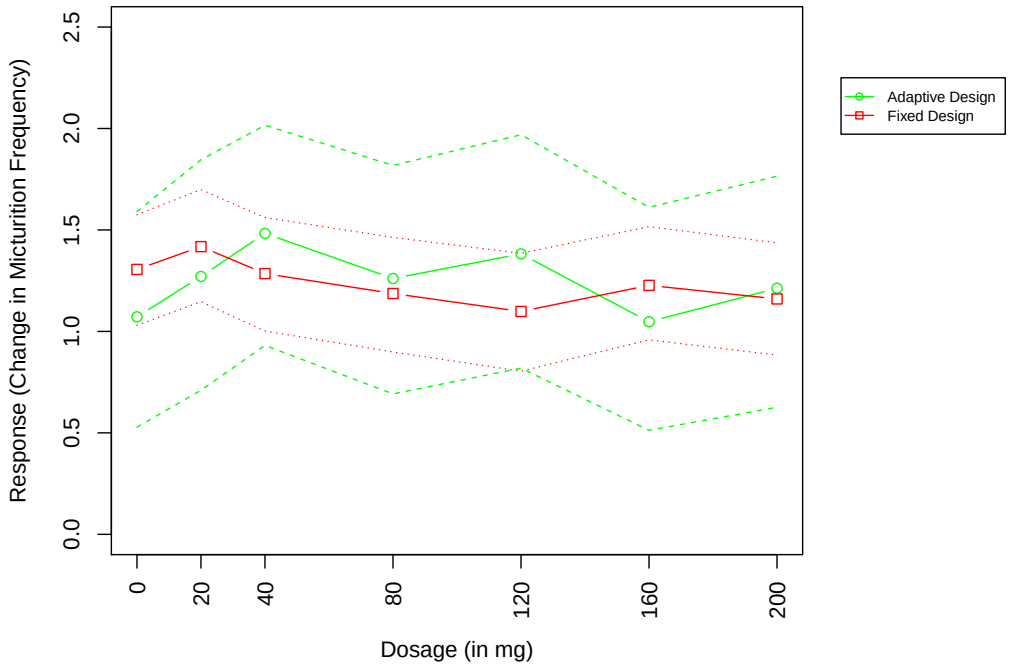

Figure 3.7: Estimated null dose-response curves with 95% credible intervals under the BADLM and the BFPADLM clinical trial designs.

We also considered an effective treatment dose-response curve. Figure 3.8 shows the comparison of two slowly increasing dose-response curves under the BADLM and BFPADLM design with 95% credible intervals for each. The two curves are similarly shaped with varying widths and degrees of smoothness for the credible intervals. The BFPADLM design has narrower 95% credible intervals than the BADLM design because it used more patients. However, the BADLM design chose correctly with considerably fewer patients.

Figure 3.9 shows two quickly increasing, estimated dose-response curves with 95% credible intervals. Similar to the null and slowly increasing dose-response curves above, the two estimated dose-response curves have the correct shape with varying levels of smoothness. Also, the BFPADLM design yielded narrower gave more narrow 95% credible intervals than the BADLM design because it used more patients. However, the adaptive design again chose correctly with fewer patients.

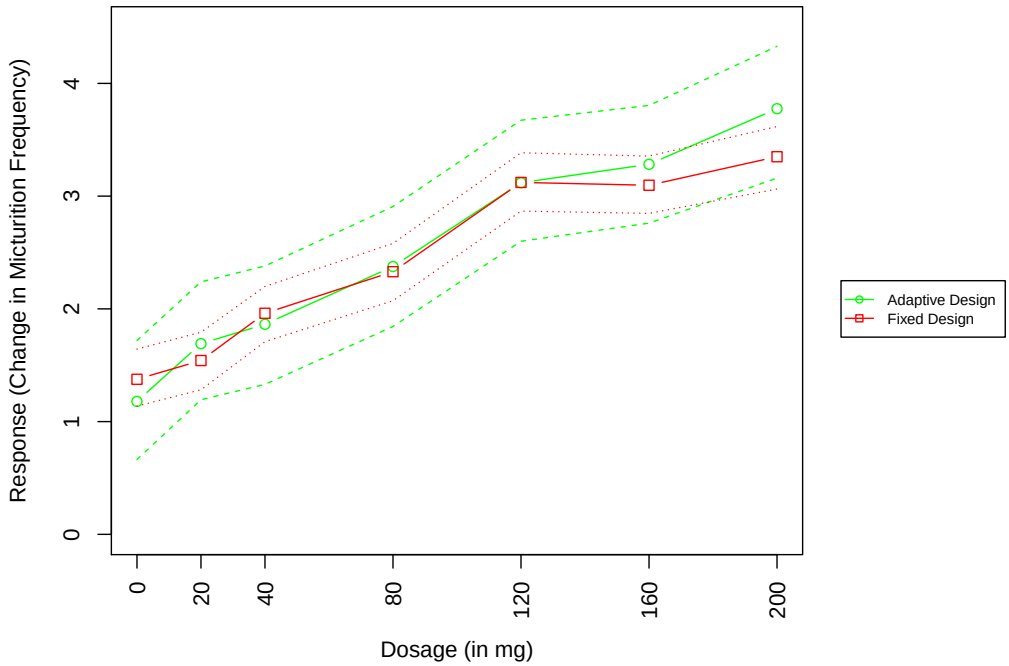

Figure 3.8: Estimated slowly increasing dose-response curves with 95% credible intervals under the BADLM and the BFPADLM clinical trial designs.

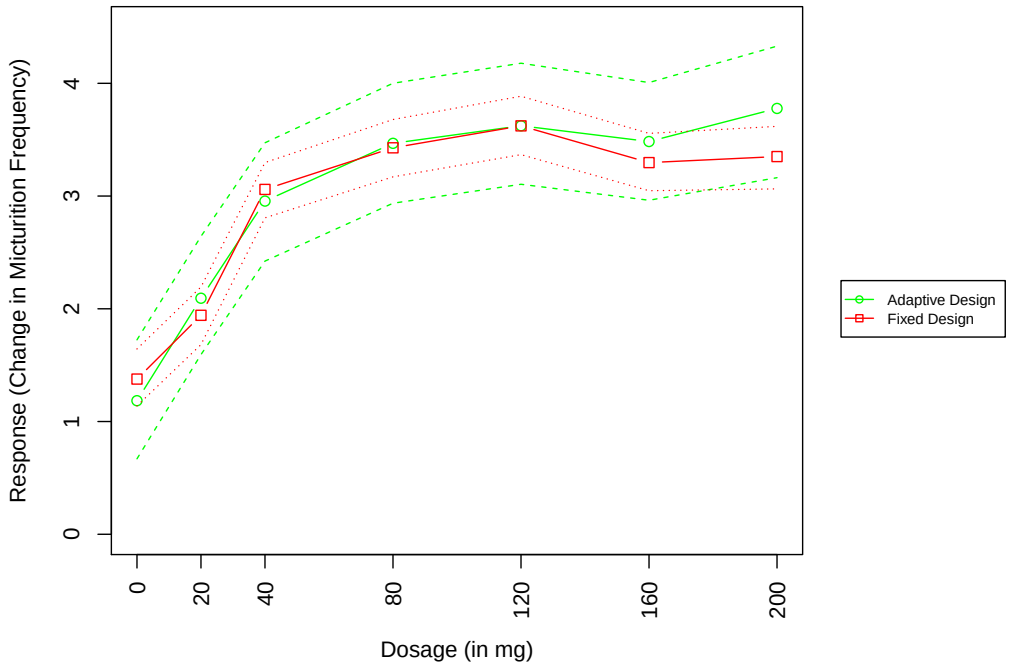

Figure 3.9: Estimated quickly increasing dose-response curves with 95% credible intervals under the BADLM and the BFPADLM clinical trial designs.

For the non-monotone dose-response curve, we see a pattern similar to the plot in Figure 3.10. The plot shows that the width of the 95% credible interval for the BFPADLM design is more narrow than that of the BADLM design. Also, the shape of the non-monotone curve is well captured under both designs. However, the BADLM design used considerably fewer patients than the BFPADLM design to model the dose-response curve correctly.

In summary, the four dose-response curves all were well estimated using both the BADLM and BFPADLM clinical trial designs, and the BFPADLM design yielded a narrower 95% credible interval while the adaptive design chose correctly with fewer patients. Therefore, our BADLM captured the four different shapes of the dose-response curves and, hence, provided a more flexible Phase II method.

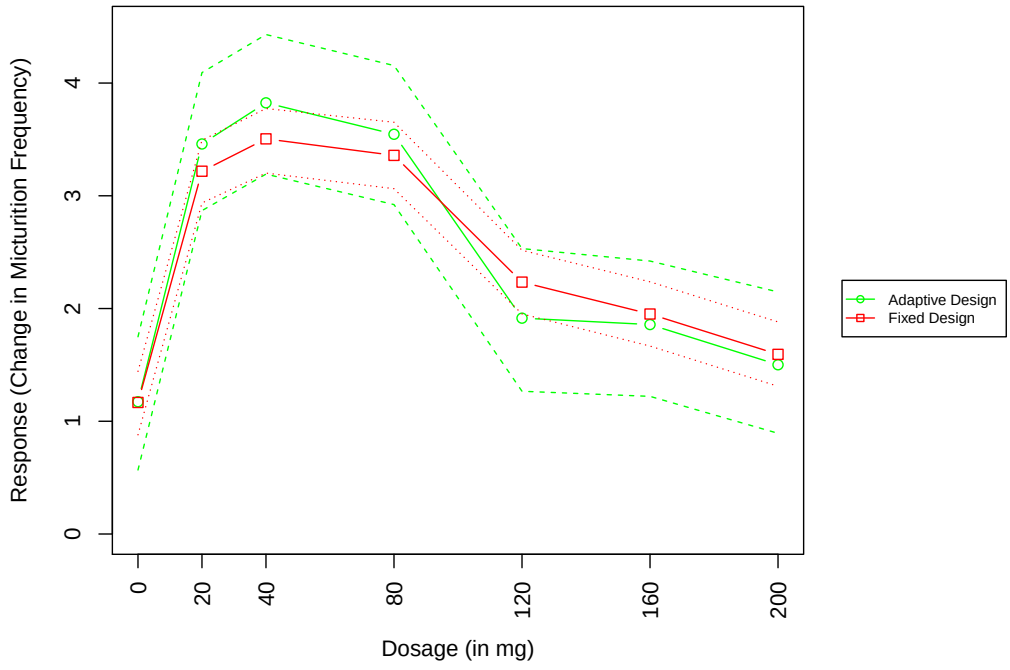

Figure 3.10: Estimated non-monotone dose-response curves with 95% credible intervals under the BADLM and the BFPADLM clinical trial designs.

3.8 Discussion

A Phase II study was performed to determine whether or not a new treatment will be further tested in Phase III with a relatively larger number of patients. The conventional design used in Phase II usually uses a logistic regression model to determine the efficacy of a new treatment with a cohort of patients that requires the response to be categorized. Also, the conventional design considers only efficacy and assumes a monotone dose-response relationship such that a higher dosage indicates a higher response. In a conventional design one simultaneously assigns an equal number of patients to each dosage without considering the dosage's effectiveness. Thus, in order to improve a Phase II trial, researchers have developed some statistical designs in the last twenty years that analyze the drug efficacy. However, many designs are for categorical data and, thus, a Phase II clinical trial design with continuous data can be difficult to utilize.

In this study, we have proposed a BADLM that is similar to a model proposed by Leininger (2010) that does not require categorical data, but bases the decision in the Phase II study on both the efficacy and the toxicity of the treatment. Instead of assuming a monotone dose-response curve for efficacy, we have considered four types of dose-response relationships that are commonly encountered in clinical trials. We have also designed a Bayesian approach for patients to be allocated to more effective treatments proportional to the efficacy and toxicity of the treatments. Our approach provides an alternative to the present Phase II clinical trial designs and is useful for the determination of drug efficacy when the response is continuous.

We have compared our proposed BADLM clinical trial design to a BFPADLM design with an equal number of patients allocated to each dosage. We have found that the bivariate normal DLM can satisfactorily estimate the four different dose-response curves. Also, the patients were used more efficiently because we assigned them to more effective dosages. Although the BFPADLM design provides a more smoothly

estimated dose-response curve with shorter credible intervals, the BADLM design used fewer patients than the BFPADLM design to reach the same decision, especially for the null dose-response relationship. For the null dose-response relationship, the adaptive design chose correctly with only 84 patients while the BFPADLM design used 350 patients. Moreover, the adaptive design chose correct decisions for the 1000 simulated trials more often than the BFPADLM design did.

We have also found that when the correlation between efficacy and toxicity was increased, the percentages of making correct decisions out of 1000 trials decreased from over 94% to over 80% for the adaptive clinical trial design. This phenomenon may have occurred because the stopping rules were determined under the small efficacy-toxicity correlation scenario so they might not be the best to use when the efficacy-toxicity correlation increases.

Our BADLM approach provides researchers with a direct way to handle continuous data rather than categorize it. Furthermore, our proposed BADLM can identify different types of dose-response curves, and, hence, provides a more flexible Phase II study.

CHAPTER FOUR

Validation of Prediction Using Logistic Regression Models in the Presence of Missing Data

4.1 Motivation

This study was motivated by the presence of missing data in logistic regression analysis. Researchers employ many methods to deal with missing data, including simple methods such as excluding all the missingness and using complicated methods such as multiple imputation. Here, we are interested in determining which missing data method that should be used when some of the variables in a data set are not fully recorded. Hence, we study various imputation methods under both missing-completely-at-random (MCAR) and missing-at-random (MAR) mechanisms for logistic regression, we validate the logistic regression prediction models, and compare imputation methods using Monte Carlo simulation.

The remainder of this chapter is organized as follows: We provide some background information concerning the research of missing data and imputation in Section 4.2. Section 4.3 introduces the imputation methods we consider, and Section 4.4.3 describes the metrics that one can use to determine the performance of each imputation method. In Sections 4.5 and 4.6, we present the design of our simulation study and the results we attained respectively. Section 4.7 concludes with a discussion of the applied methods.

4.2 Previous Logistic Regression Studies

Ambler, Omar, and Royston (2007) have investigated various imputation methods to evaluate their effect on risk-model estimation and on prediction accuracy. Specifically, they investigated Hotdeck and multiple imputation by chained equations (MICE), along with several single imputation methods. Their results suggest that

the prediction with complete-data-only analysis was poor and should be avoided; mean/mode imputation outperformed the complete-data-only method but performed worse than other considered methods; conditional-mean imputation outperformed the mean/mode method; and hotdeck imputation was the worst considered MI method and performed worse than the single imputation methods. MICE performed the best among all the imputation methods investigated. Hence, MICE was recommended for use in practice.

Steyerberg, Vickers, Cook, Gerds, Gonen, Obuchowski, Pencina, and Kattan (2010) have aimed to define the role of a list of traditional measures, such as the Brier score, concordance statistic, receiver operating characteristic (ROC) curve, and several new measures, including net reclassification improvement (NRI) and integrated discrimination improvement (IDI). For illustration, they have presented a case study that predicts the presence of residual tumor versus benign tissue in patients with testicular cancer. They have also suggested that one should always report discrimination and calibration for a prediction model and that decision-analytic measures should be reported if the predictive model is to be used for clinical decisions.

4.3 Imputation Methods and Logistic Regression

4.3.1 Imputation Methods

Complete-data-only Analysis The complete-data-only approach to the missing-data problem is to remove all observations with missing elements. That is, we exclude all the units for which the outcome of any of the covariates is missing. However, two problems arise with this approach (Gelman and Hill, 2006):

1) This action could bias the analysis if the units with missing values differ systematically from the completely observed cases, or

2) Most of the data could be excluded for the sake of a simple statistical analysis if many variables are initially included in a model.

Single-Mean Imputation Rather than excluding observations with missing values, one can use single-mean imputation, which is one of the simplest approaches to imputing the missing values in a dataset. The advantage of this method is that all information is kept in the data. However, by employing single-value imputation, we are essentially pretending that the imputed missing values are known with certainty. Thus, single-mean imputation can yield a different type of bias from that of the complete-data-only analysis.

Conditional-Mean Imputation Conditional-mean imputation is a model-based imputation method. Here, we apply the univariate regression model iteratively to the variables with missing values in the data with the following steps:

1. Obtain preliminary imputations for each predictor.
2. Fit the regression model relating the predictor with missing values to all other predictors.
3. Use least squares/logistic regression for continuous/categorical predictors.
4. Iteration is required when several predictors have missing values.
5. Stop the procedure when the regression coefficients become stable to four decimal places.

The regression model allows interactions to be considered if necessary, and the set of separate regression models makes this method easy to understand. Thus, people should be able to fit a reasonable model at each step.

On the other hand, Gelman (2006) suggested caution to ensure that the separate regression models are consistent with each other when one is using an iterative approach. For instance, imputing blood pressure on age but then ignoring blood pressure when imputing age would be impractical.

Multiple Imputation by Chained Equations Instead of imputing with a single value, multiple imputation creates more than one (say, 5) imputed value for each missing value, thus creating multiple data sets. A standard analysis can be performed on each

completed dataset, and then the inferences are combined across datasets. Multiple imputation reflects uncertainty about the imputation model, which is an advantage compared to the methods described above.

The process of Multiple Imputation by Chained Equations (MICE) is conducted as follows:

1. Obtain preliminary imputations for each predictor.
2. Fit a regression model relating the predictor containing missing values to all other predictors.
3. Use least squares/logistic regression for a continuous/categorical predictor.
4. Estimate coefficients and covariance matrices.
5. Generate coefficients from a multivariate normal distribution.
6. Impute the missing data.
7. Cycle over all predictors with missing values.

Hotdeck Imputation Hotdeck imputation employs observed values from the same data set to replace the missing points. It can be viewed as a nonparametric imputation method. Two basic steps are involved in Hotdeck imputation:

1. Draw n_{miss} with replacement from the n_{obs} observed data, where n_{miss} is the number of missing values and n_{obs} is the number of observed values
2. Draw n_{miss} values with replacement from these n_{miss} values to use as imputed values.

4.3.2 Logistic Regression

At times the dependent variable of interest is binary. For example, a person is male or female; a person votes in the presidential election or does not; a home loan was paid back or was not. Additionally, we have independent variables that are either discrete or continuous, which motivates the problem of identifying which category a new observation belongs to. The basic idea of statistical classification is to determine the membership of a new observation based on the data in which each observation is

truly categorized. To do so, we consider probabilities, and, hence, we need a model (Agresti, 2012).

Assume we have a dependent variable Y with two outcomes, 1 and 0, and an independent variable X. We want to model the conditional probability $\Pr(Y = 0|X = x)$ as a linear function of x. However, we must have $0 \leq \Pr(Y = 0|X = x) \leq 1$ while a linear function of x is not bounded. To address this issue, we first transform $\Pr(Y = 0|X = x)$ so that it is unbounded and then model it as a linear function of x. One such transformation is called the logit function, where

$$\text{logit}\Pr(Y = 0|X = x) = \log\left(\frac{\Pr(Y=0|X=x)}{1-\Pr(Y=0|X=x)}\right).$$

Modeling the logistic transformation of $\Pr(Y = 0|X = x)$ as a linear function of x is called logistic regression. The form of the logistic regression model is

$$\log\left(\frac{\Pr(Y=0|X=x)}{1-\Pr(Y=0|X=x)}\right) = \beta_0 + \beta_1 x,$$

and, therefore,

$$\Pr(Y = 0|X = x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}.$$

We predict $Y = 0$ if $\Pr(Y = 0|X = x) \geq 0.5$ and $Y = 1$ if $\Pr(Y = 0|X = x) \leq 0.5$.

Logistic regression is widely used for classification of multivariate unlabeled data. We usually estimate the coefficients in the model by using maximum likelihood. For one data point (x_i, y_i), the likelihood is

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}.$$

However, one cannot derive a closed-form expression for the coefficient that maximizes the likelihood function as one can with linear regression. Instead, we can obtain MLEs numerically with iterative processes, such as Newton's (Newton-Raphson) method (Kaw and Kalu, 2011).

The interpretation of the logistic regression coefficients is also different from a simple linear regression. The coefficients of the logistic model describe the change

in the probability of observing $Y = 0$ expressed on the logistic scale instead of the change in the Y value directly, as in the simple linear regression.

The dependent variable Y can take on more than two categories, and logistic regression will also be suitable with more than two parameters, β_0 and β_1. The parameter estimation proceeds as before with maximum likelihood (Agresti, 2012). Our study is concentrated on binary discrete dependent variables.

4.4 Performance Measurements

4.4.1 Prediction Measurement

Mean absolute error (MAE) measures how close predictions are to actual outcomes. The MAE is

$$\text{MAE} \equiv \frac{1}{n} \sum_{i=1}^{n} |f_i - y_i|,$$

where f_i is the i^{th} prediction value, y_i is the i^{th} true value, n is the total number of observations, and $i = 1, \ldots, n$.

4.4.2 Discrimination

Discrimination is a measure of how well the subjects with and without an outcome of interest are separated by a logistic regression model. For binary outcomes, the area under the receiver operating characteristic (ROC) curve, which is a plot of the sensitivity (the true positive rate) against 1-specificity (false positive rate) for consecutive cutoffs for the probability of an outcome, is the AUC. The AUC is the most commonly used measurement for discrimination of a binary regression model.

An alternative measure to the full AUC is the partial AUC, which is the area under the ROC curve where data have been observed. Walter (2005) has suggested several disadvantages of the partial AUC measurement. In contrast to the full AUC, the partial AUC is not as robust to heterogeneity and thus can obscure the result of comparisons between tests.

4.4.3 A Goodness-of-Fit Statistic

The Hosmer-Lemeshow statistic measures the goodness-of-fit for logistic regression models. The hypotheses of the test statistic are as follows:

H_0: the model fits well.

H_A: the model does not fit well.

The Hosmer-Lemeshow statistic is

$$H \equiv \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i(1 - E_i/n_i)} \sim \chi^2_{n-2},$$

where

$$n \equiv \text{Number of groups},$$

$$n_i \equiv \text{Number of observations in the } i^{th}\text{group},$$

$$O_i \equiv \text{Observed number of cases in the } i^{th}\text{group},$$

and

$$E_i = \text{Expected number of cases in the } i^{th}\text{group}.$$

4.5 A Monte Carlo Simulation for Imputation Efficacy in Logistic Regression

4.5.1 Data Simulation

The Monte Carlo simulation is based on the testicular cancer dataset from Steyerberg et al. (2010). The training dataset (544 patients) includes five predictors for model development, and the test dataset (273 patients) with the same five predictors is for the evaluation of the developed logistic regression prediction model. Table 4.1 gives the data description.

Table 4.1: Description of testicular cancer dataset.

Name	Description (no/yes is coded as 0/1)
nec	Necrosis at resection (0-1)
preafp	Prechemotherapy AFP normal? (0-1)
prehcg	Prechemotherapy HCG normal? (0-1)
lnldhst	Ln of standardised prechemotherapy LDH
sqpost	Square root of postchemotherapy mass size
reduc10	Reduction in mass size per 10%: (pre-pos)/pre*10

Instead of using the real data, we used the testicular cancer data to parameterize the simulation data, and we used R for our simulation software. For continuous variables, we generate data from a Beta distribution estimated from the real data. For categorical variables, 0/1 (no/yes) were generated from a Bernoulli distribution with probability of success p estimated from the real data. For continuous variables, the data was shifted and transformed so that the data has a similar mean and covariance structure to that of the real data. The estimated Beta distributions for the modeled data from that we sampled are shown in Figure 4.1.

To create a dataset with MCAR missing values, we first generated 0/1 (no/yes) data from a Bernoulli(p) distribution with $p = 0.1, 0.2, 0.3$. Next, we formed a matrix with the generated binary data with the same dimension as the full data. Last, we performed a Hadamard product of the formed matrix of 0's and 1's and the full data matrix while the outcome was excluded. The missingness exists only in the predictor variables, and none of the outcome data was missing.

To simulate an MAR dataset, we made the initial process the same as that for MCAR data. According to the definition of MAR, the probability of an observation being missing depends only on the observed variables. Two of the five variables are fully recorded while the remaining three variables were simulated with missingness

where the probability of missingness depended on the two fully-recorded variables (Ambler and Omar, 2007).

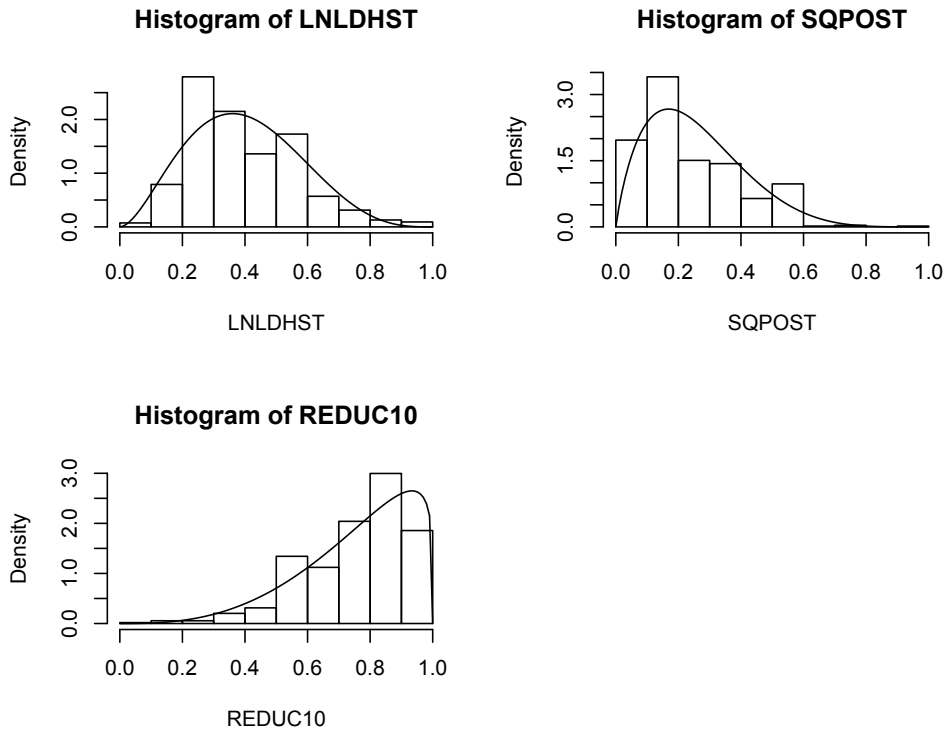

Figure 4.1: A Histogram of continuous variables with beta distributions fitted.

4.5.2 Simulation Design

The training-data sample sizes considered here were 30, 50, and 100. The test-data size was 1000. Three levels of POM were used: 0.1, 0.2, and 0.3. We compared the following imputation methods for each dataset:

- Complete-data-only analysis
- Single-mean imputation
- Conditional-mean imputation
- MICE ($m = 5$)
- Hotdeck ($m = 5$)

where m is the number of imputed values for each missing value. After the imputation, we then fit a logistic regression model to the completed training data and estimated the covariate coefficients. We then predicted the outcomes using the test data, the estimated logistic model, and

- MAE
- AUC
- HL p-value,

which are measurements for prediction comparison.

We set the number of iterations to be $k = 1000$. We also conducted the logistic regression prediction with the full training data (no missingness) for comparison purposes. In addition, we remark that the simulations were conducted for both MCAR and MAR mechanisms.

4.6 The Monte Carlo Simulation Results

4.6.1 Missing-Completely-at-Random (MCAR)

4.6.1.1 *Prediction measurement.* In the simulation we had three levels of percent of missingness (POM)—0.1, 0.2, 0.3—for each training sample size, $N = 30,\ 50,\ 100$. The MAE was computed for each combination of POM and sample size. Note that the complete-data-only analysis was not conducted for $N = 30$ with POM $= 0.3$ and $N = 50$ with POM $= 0.3$ because a high POM with small sample sizes causes exclusion of most of the data and, hence, a singularity problem. A small MAE suggests a small error rate and, thus, better prediction. In Figures 4.2-4.4 we display a boxplot of the MAE for each sample size with three levels of POM.

The prediction MAE for the full data was 0.2485. Compared with the other methods, for the complete-data-only analysis, we had MAE $= 0.3072$ for $N = 30$ with

POM = 0.1 and MAE = 0.4031 for $N = 30$ with POM = 0.2. These results indicated poor prediction properties. Moreover, the complete-data-only MAE was greater than the MAE for the other imputation method as we can see from Figure 4.2. We can also see that the MAE for the single-value and multiple imputation methods all increased as the level of POM increased. The conditional-mean imputation had the highest MAE when the POM = 0.3. Also, the MAE of single-mean and hotdeck imputation was close to that of the full data. The single-mean imputation method performed surprisingly well since it is generally considered the worst imputation method the researcher should use.

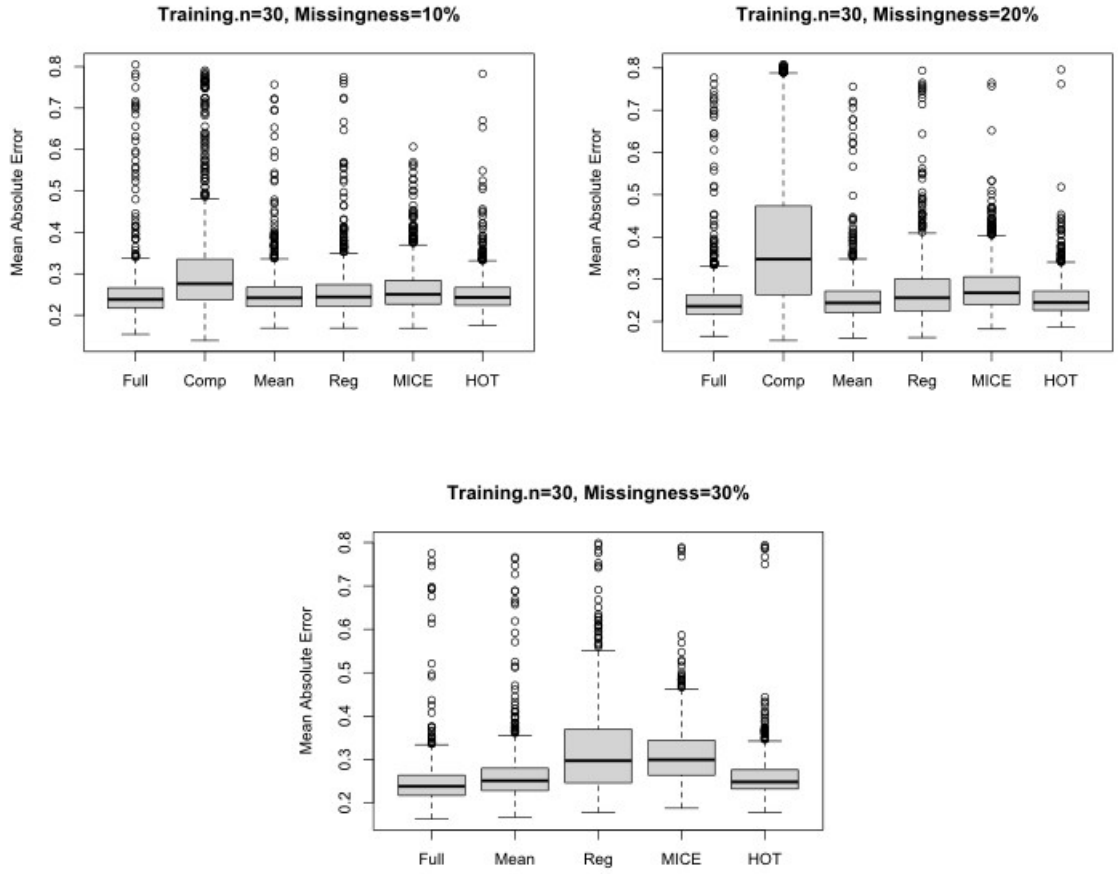

Figure 4.2: Boxplots of the MAE for five imputation methods and for the full data with $N = 30$, POM $= 0.1, 0.2, 0.3$, when the data is MCAR.

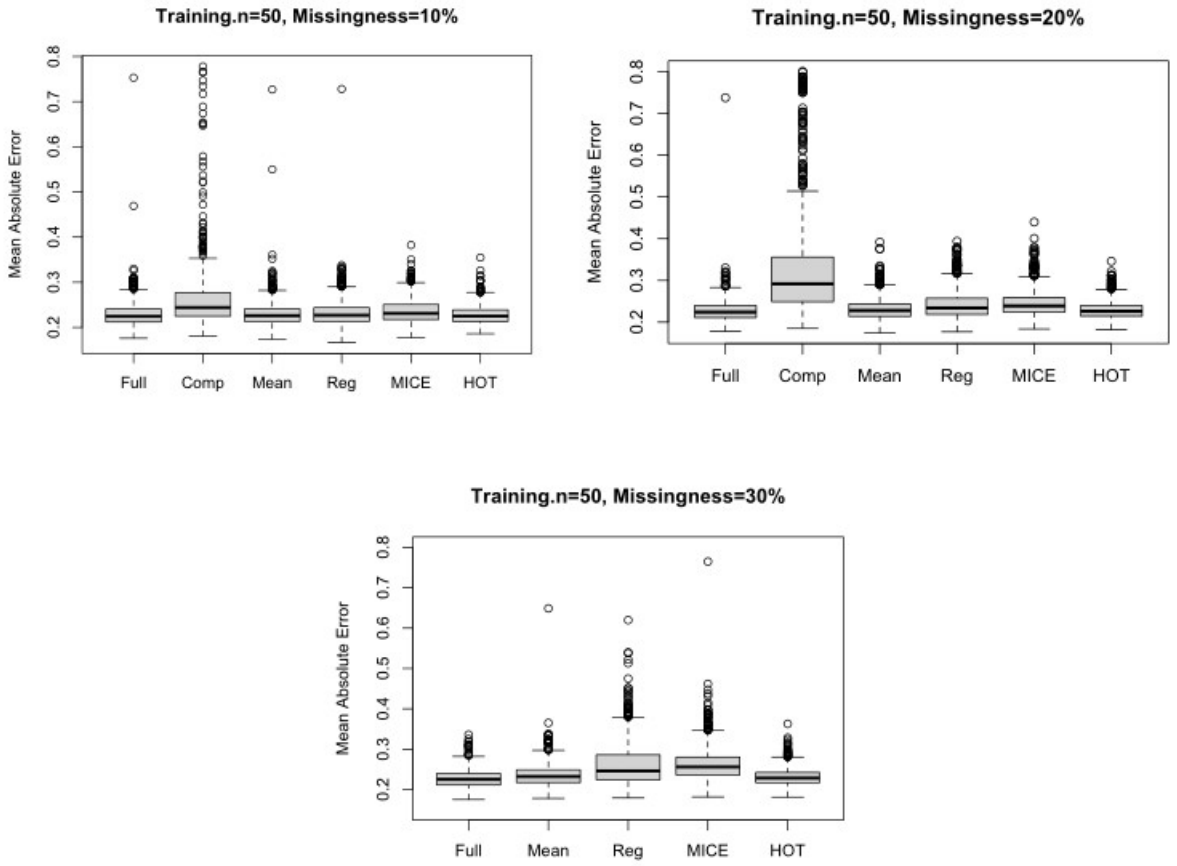

Figure 4.3: Boxplots of the MAE for five imputation methods and for the full data with $N = 50$, POM $= 0.1, 0.2, 0.3$, when the data is MCAR.

Figure 4.3 displays similar results for $N = 50$ with the three levels of POM. Compared to the case when $N = 30$, the MAE for all five imputation methods and the full data slightly decreased because of the larger sample size. The complete-data-only analysis had a larger MAE than did the other imputation methods. The MAE for the hotdeck and single-mean imputation methods was less than the other considered imputation methods for POM = 0.1 and POM = 0.2. Also, the differences between the MAE for the hotdeck and the competing imputation methods were even greater for the POM = 0.3. Surprisingly, the single-mean imputation performed approximately as well as the hotdeck imputation method when $N = 30$.

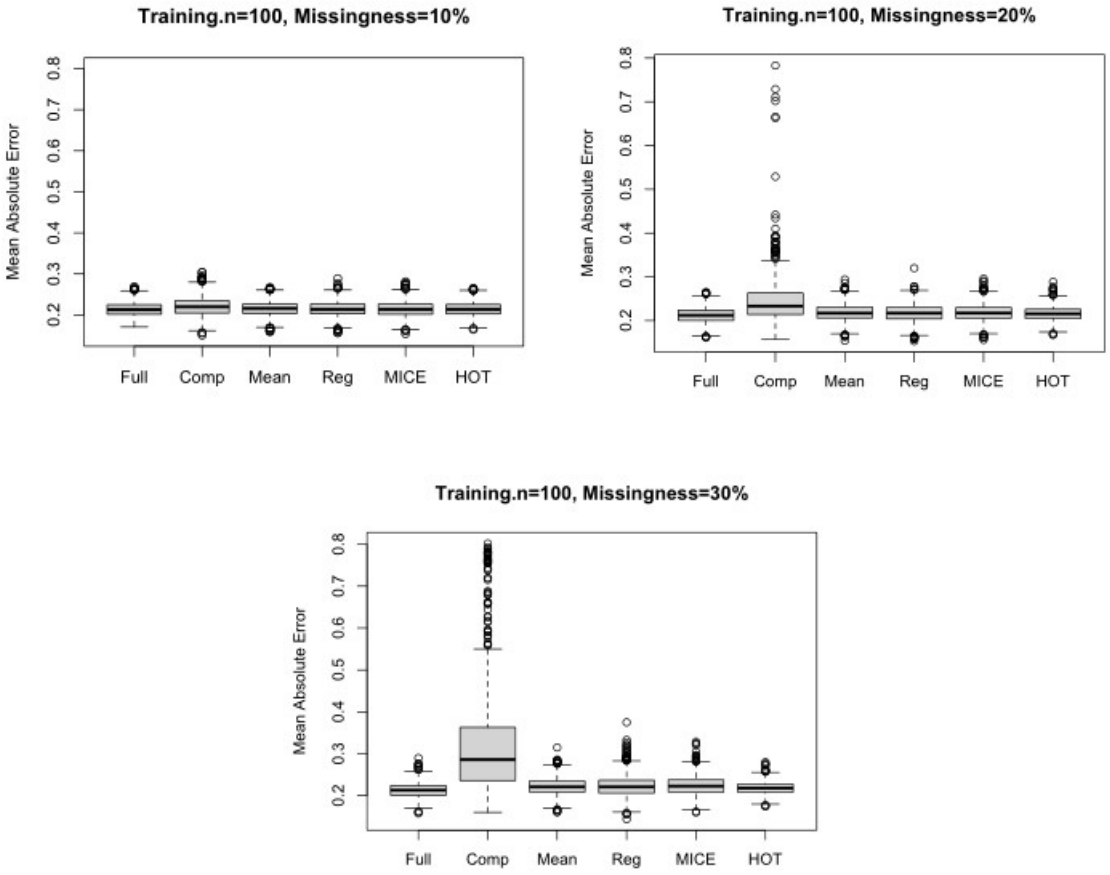

Figure 4.4: Boxplots of the MAE for five imputation methods and for the full data with $N = 100$, POM $= 0.1, 0.2, 0.3$, when the data is MCAR.

When the training-sample size was increased to $N = 100$, the MAEs were very similar aside from the complete-data-only analysis. The complete-data-only analysis

gave an MAE = 0.2204, which was close to the full-data prediction (MAE = 0.2125) when the POM = 0.1 and increased to MAE = 0.2438 and MAE = 0.3226 with POM = 0.2 and POM = 0.3, respectively. Thus, the complete-data-only imputation method yielded very poor prediction results for higher levels of POM. The MAE for the other imputation methods increased as the level of POM increased but was still close to the full data MAE. The hotdeck imputation method had the smallest MAE indicating the best performance among all considered imputation methods in terms of MAE.

4.6.1.2 *Discrimination utility.* In this section, we compare the discrimination utility of the five imputation methods for logistic regression using the AUC as our criterion for prediction ability.

The AUC was reported for $N = 30,\ 50,\ 100$, with levels of POM = 0.1, 0.2, and POM = 0.3 in Figures 4.5-4.7. For $N = 30$, the AUCs were all above 0.5, and the complete-data-only analysis had a smaller AUC than the other four imputation methods, thus indicating that the other imputation methods yielded superior discrimination. The AUCs of the single-mean imputation, MICE, and hotdeck imputation were close to that of the full data when POM = 0.1, 0.2. When we increased to POM = 0.3, the single-mean imputation procedure was the closest to the full data in terms of AUC. Also, the AUC for the conditional-mean regression method decreased slightly.

For $N = 50$, the results were similar to the results for $N = 30$. The complete-data-only analysis had the smallest AUC for $N = 30$ and 50, while the AUCs of the other four imputation methods were close to the AUC for the full data prediction. For the POM = 0.3, the hotdeck imputation yielded the smallest AUC than the other considered imputation methods.

Figure 4.7 shows the AUC value for $N = 100$. The AUCs of all considered imputation methods were close to the full-data AUC with POM = 0.1, 0.2. The

complete-data-only AUC and hotdeck imputation AUC were less than the other considered imputation methods. All of the AUCs were close to or above 0.6 for $N = 100$ with the complete-data-only having the smallest AUC when POM = 0.3.

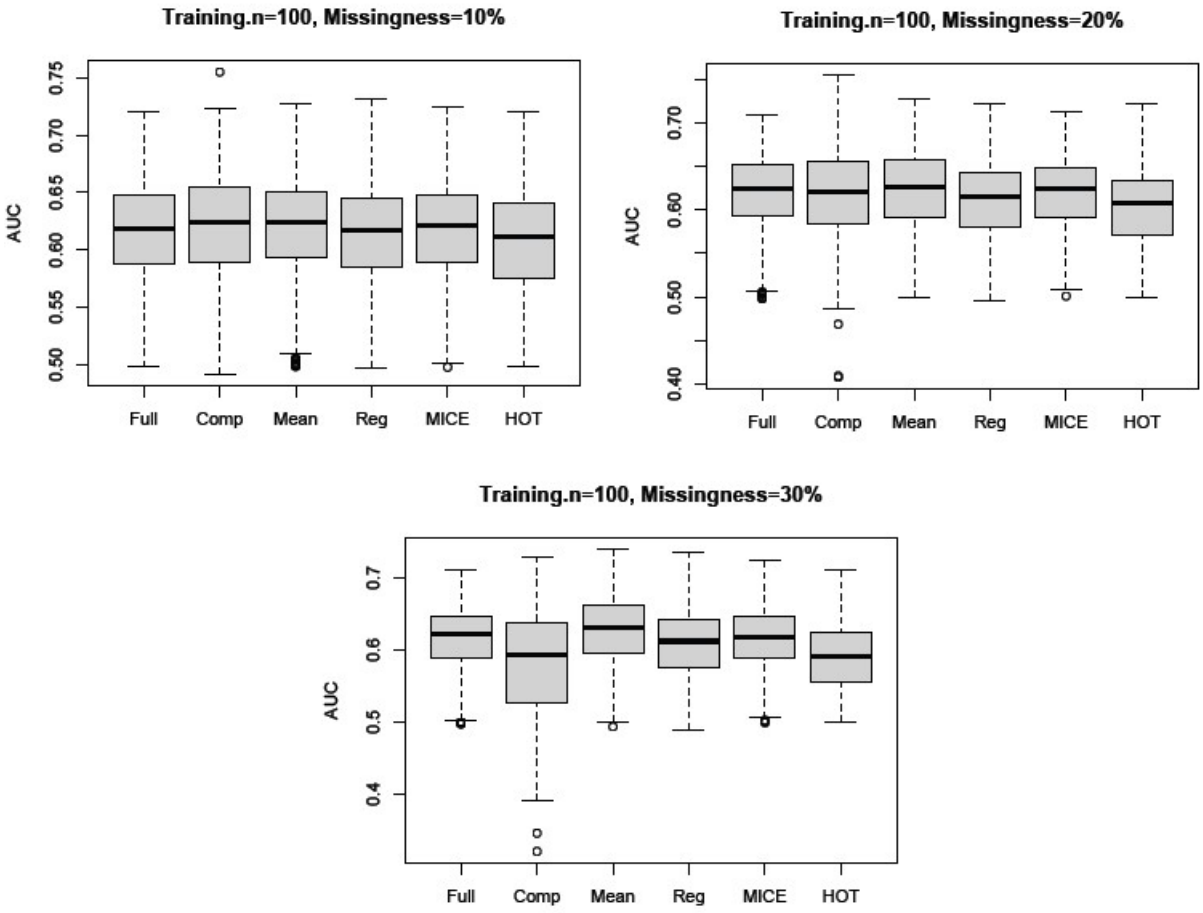

Figure 4.7: Boxplots of the AUC for five imputation methods and for the full data with $N = 100$, POM $= 0.1, 0.2, 0.3$, when the data is MCAR.

4.6.1.3 *A Goodness-of-Fit statistic.* Figures 4.8-4.10 show the Hosmer Lemeshow statistic results in terms of the p-value. For $N = 30$ with POM = 0.1, 0.2, 0.3, all p-values were above 0.6; for $N = 50$ and 100, with the three levels of POM, all p-values were above 0.5. These results indicated that the prediction model fit well for each imputation method. However, we remark that a well-fitting model does not imply a model that predicts well.

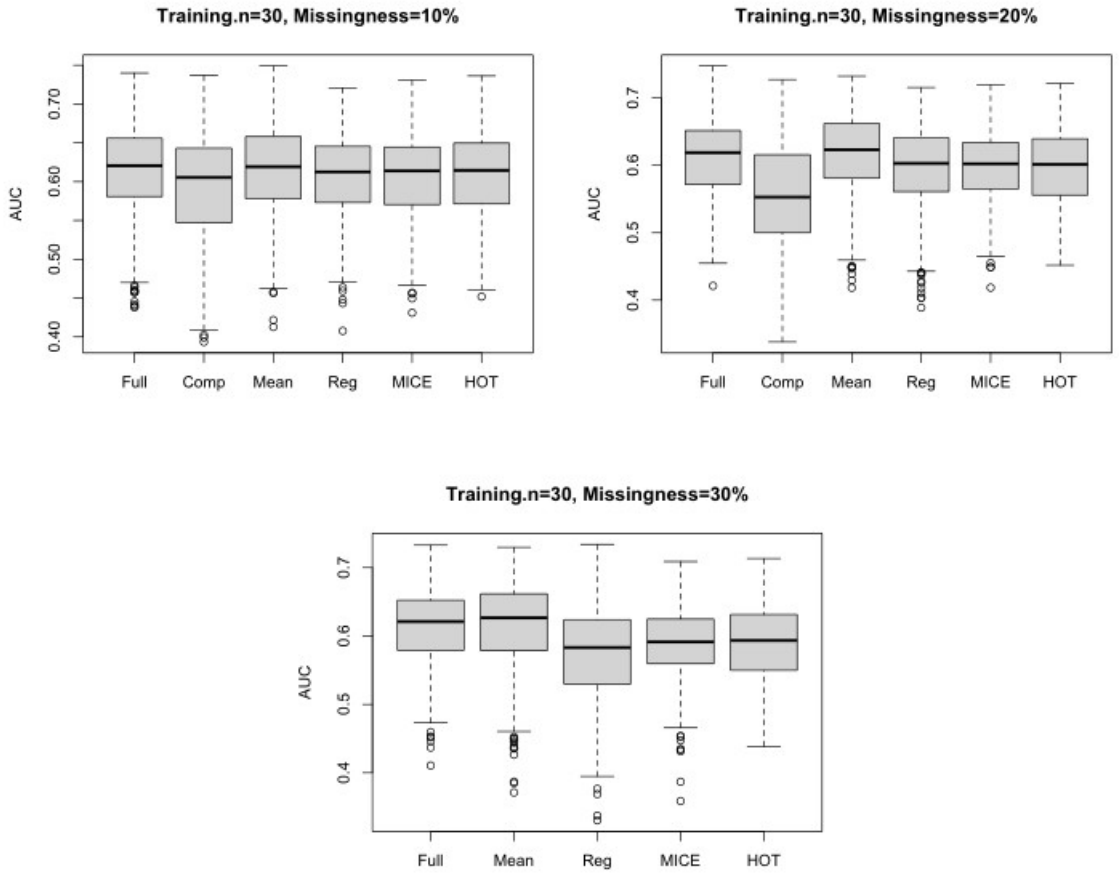

Figure 4.5: Boxplots of the AUC for five imputation methods and for the full data with $N = 30$, POM $= 0.1, 0.2, 0.3$, when the data is MCAR.

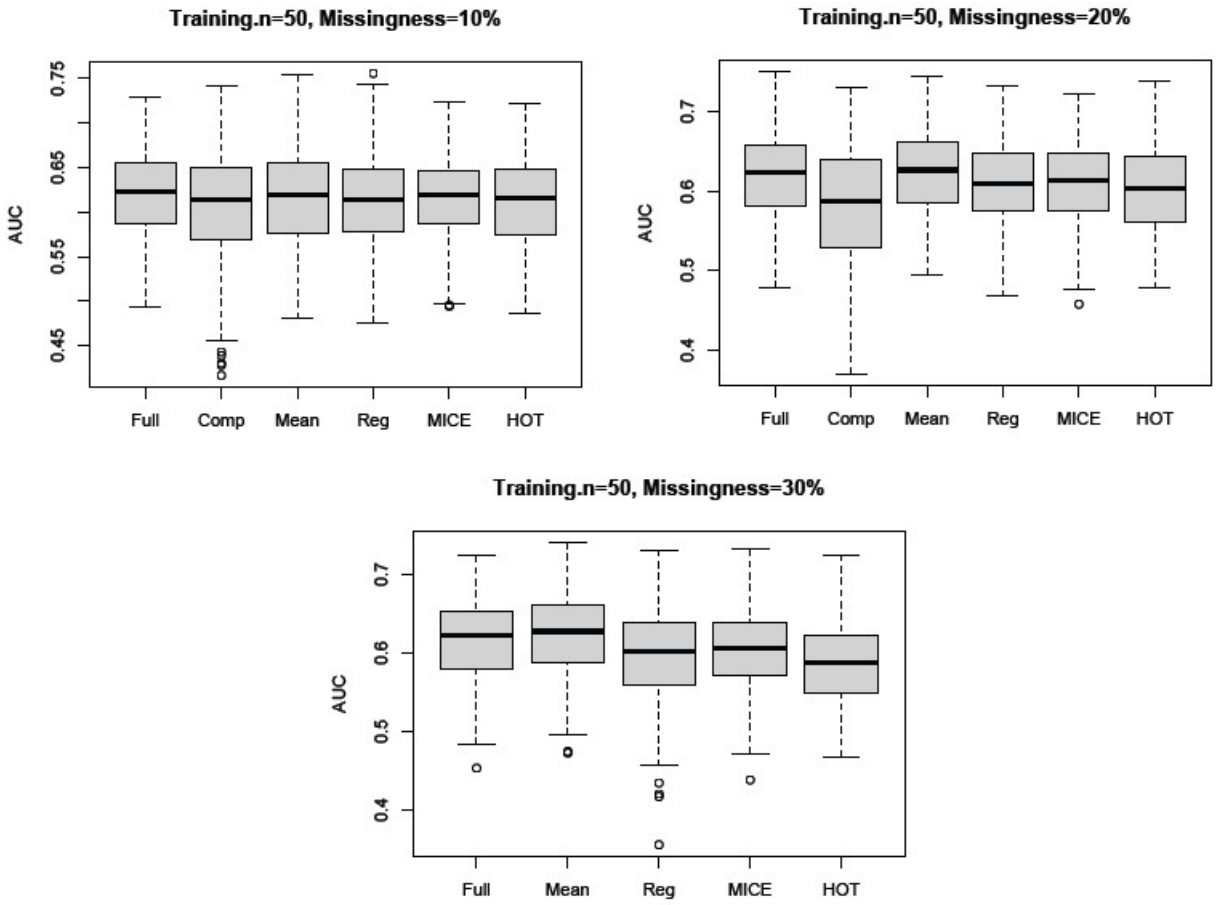

Figure 4.6: Boxplots of the AUC for five imputation methods and for the full data with $N = 50$, POM $= 0.1, 0.2, 0.3$, when the data is MCAR.

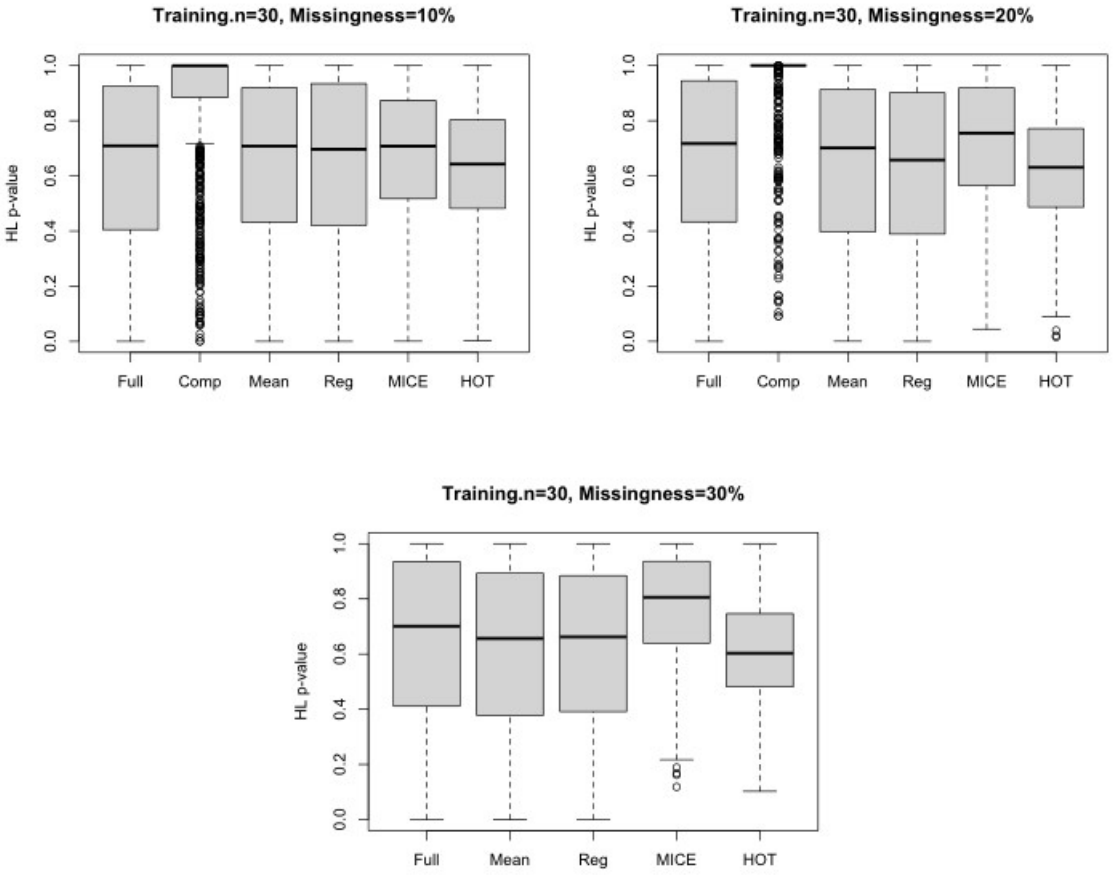

Figure 4.8: Boxplots of the Hosmer-Lemeshow p-value for five imputation methods and for the full data with $N = 30$, POM $= 0.1, 0.2, 0.3$, when data is MCAR.

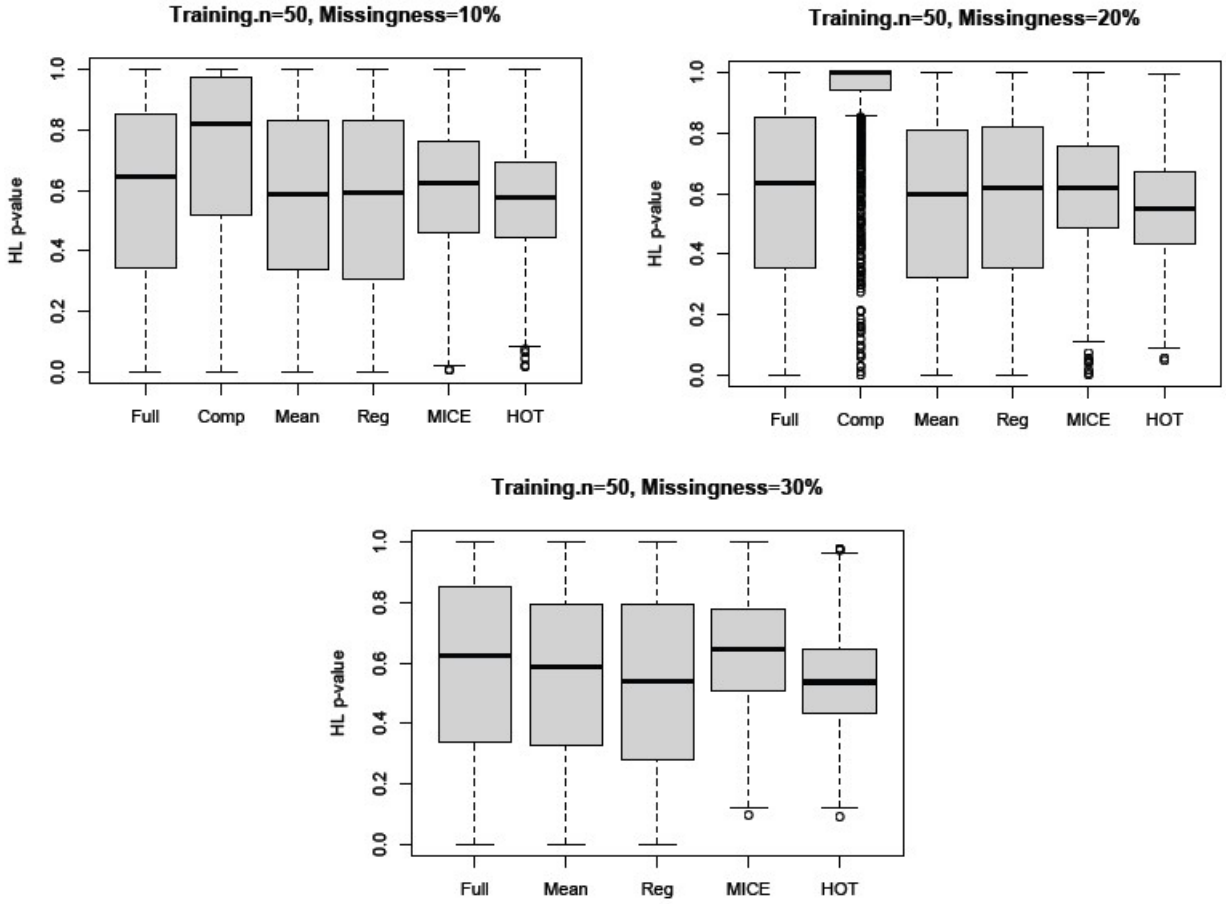

Figure 4.9: Boxplots of the Hosmer-Lemeshow p-value for five imputation methods and for the full data with $N = 50,\ \text{POM} = 0.1, 0.2, 0.3$, when the data is MCAR.

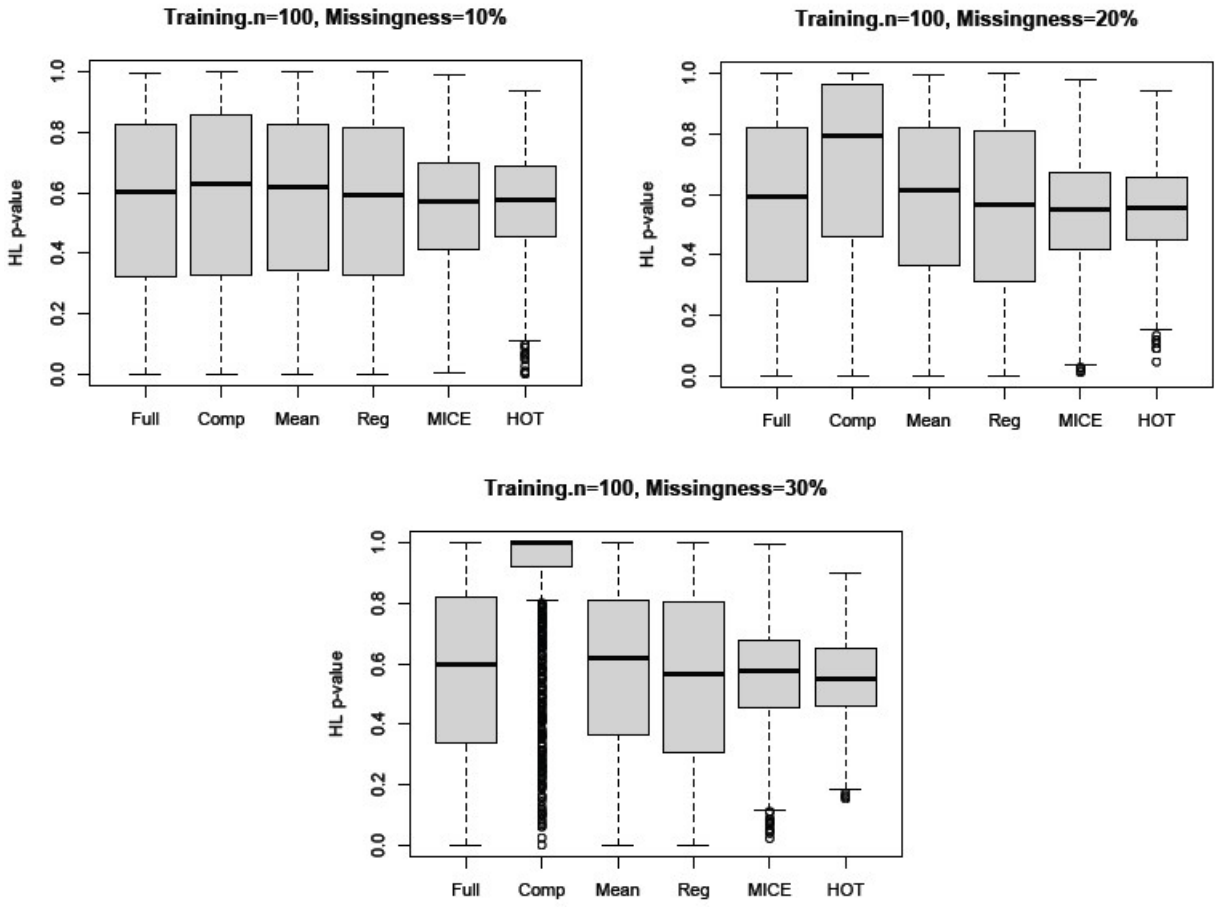

Figure 4.10: Boxplots of the Hosmer-Lemeshow p-values for five imputation methods and for the full data with $N = 100,\ \text{POM} = 0.1, 0.2, 0.3$, when the data is MCAR.

4.6.2 Missing-at-Random (MAR)

The analysis performed on MCAR data was also applied to MAR data. The simulation results with the MAR mechanism were similar to that of MCAR in the previous section. The levels of POM and sample sizes were also set to be POM = 0.1, 0.2, 0.3, and $N = 30,\ 50,\ 100$, respectively.

4.6.2.1 *Prediction measurement.* Figures 4.11-4.13 show the boxplots of the MAE for MAR data for each considered sample size with the three considered levels of POM. Also, the complete-data-only analysis was not conducted for $N = 30$ with POM = 0.3 and $N = 50$ due to a singularity problem.

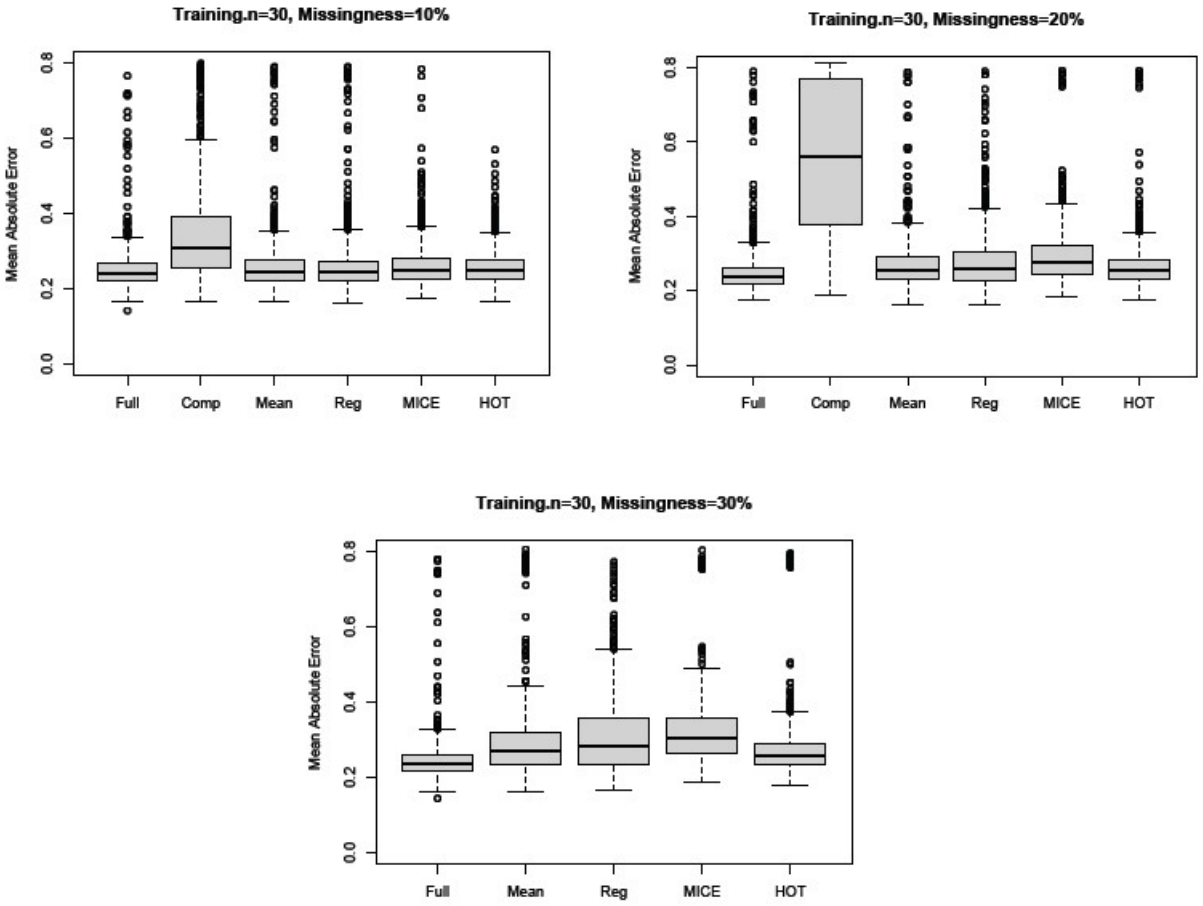

Figure 4.11: Boxplots of the MAE for five imputation methods and for the full data with $N = 30$, POM $= 0.1, 0.2, 0.3$, when the data is MAR.

The results in Figure 4.11 are similar to the results in Figure 4.2. We remark that complete-data-only analysis is inferior to the considered imputation methods because for $N = 30$, the MAEs for the complete-data-only analysis with POM $= 0.1$, 0.2 are MAE $= 0.3445$ and MAE $= 0.5589$, respectively, which are greater than the MAEs for the other imputation methods. The MAEs of all considered imputation methods increased when we increased the sample size from $N = 30$ to $N = 100$. The single-mean and hotdeck imputation method outperformed the competing imputation methods for a large sample size.

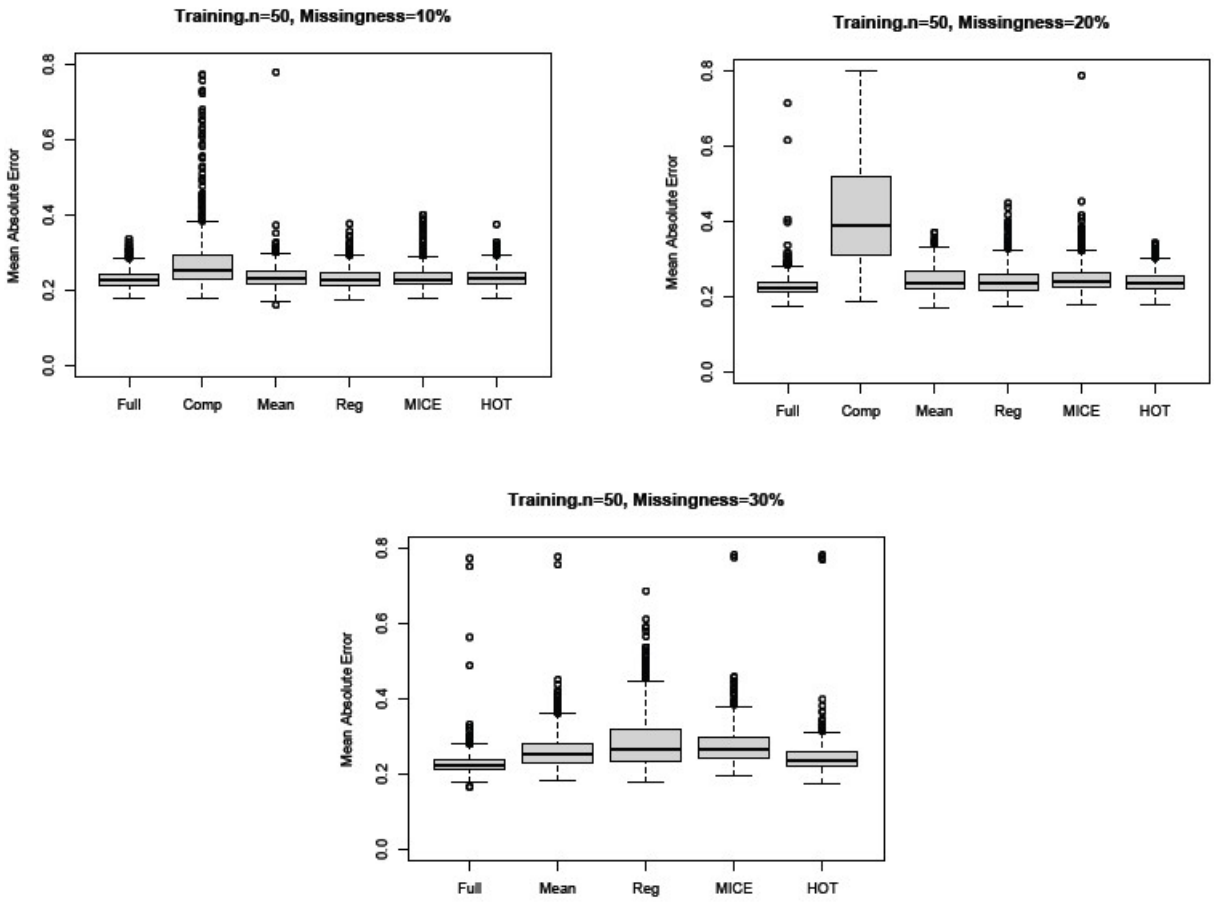

Figure 4.12: Boxplots of the MAE for five imputation methods and for the full data with $N = 50$, POM $= 0.1, 0.2, 0.3$, when the data is MAR.

For $N = 50$, the complete-data-only analysis yielded the largest MAE for POM $= 0.1$, 0.2, indicating that it had the worst prediction performance among the considered imputation methods. Excluding the complete-data-only imputation method, the MAEs of the remaining methods differed only slightly until the single-mean and hotdeck imputation began to perform better with a slightly lower MAE with POM $= 0.3$.

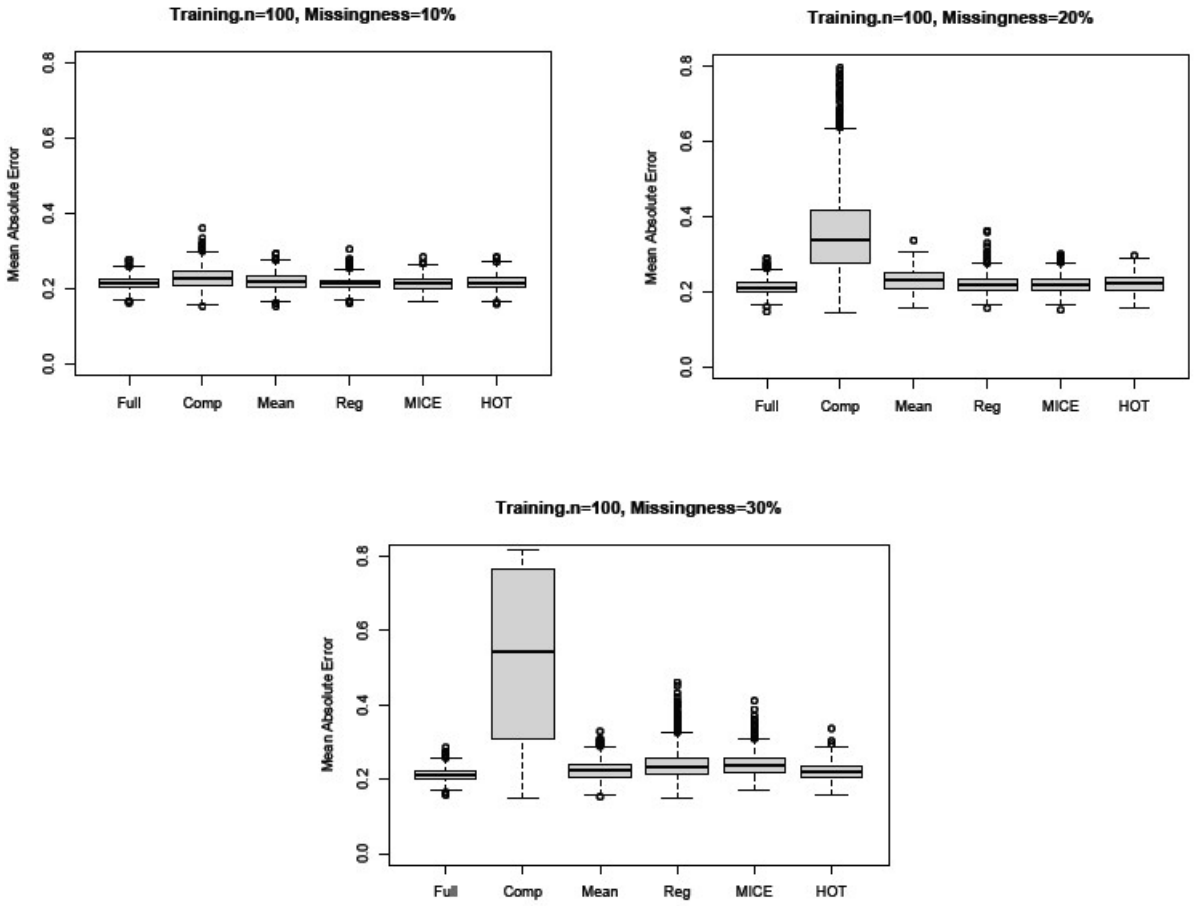

Figure 4.13: Boxplots of the MAE for five imputation methods and for the full data with $N = 100,\ \text{POM} = 0.1, 0.2, 0.3$, when the data is MAR.

The MAEs of all considered imputation methods were close to one another when we increased the sample size to $N = 100$ except for the complete-data-only analysis. The MAEs of complete-data-only analysis increased from 0.2280 to 0.3609 for an increase of POM = 0.1 to POM = 0.2, and to 0.5303 for a POM = 0.3. Similar to the results shown in Figure 4.4, the MAE of the other imputations slightly increased as the level of POM increased but still remained close to the full data, and the hotdeck imputation performed the best with the smallest MAE.

4.6.2.2 *Discrimination.* Here, the AUC was also reported for the MAR mechanism with $N = 30,\ 50,\ 100$, and the three levels of POM provided in Figures 4.14-4.16.

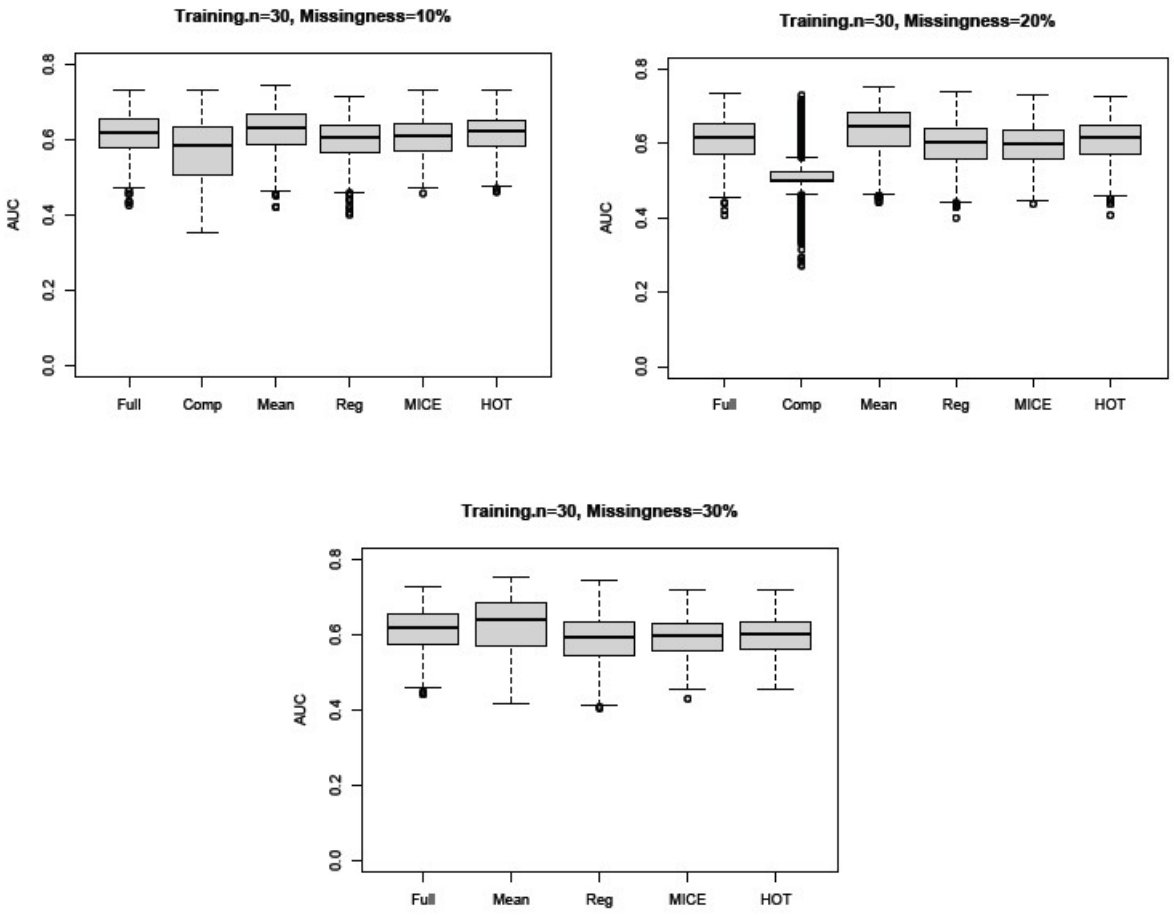

Figure 4.14: Boxplots of the AUC for five imputation methods and for the full data with $N = 30,\ \text{POM} = 0.1, 0.2, 0.3$, when the data is MAR.

From Figure 4.14 we can see that, for $N = 30,\ 50$, the AUCs were all greater than 0.5, and the complete-data-only analysis yielded the smallest AUC, which implies that it did not perform well in terms of discrimination. As the POM approached POM = 0.3, the AUC of the conditional-mean regression yielded the minimum AUC.

As in Figure 4.6, Figure 4.15 implies that the complete-data-only analysis yielded the lowest AUC for $N = 30,\ 50$, while the other considered methods yielded AUCs that were close to the full data prediction AUC. However, the conditional mean regression yielded the lowest AUC when POM = 0.3.

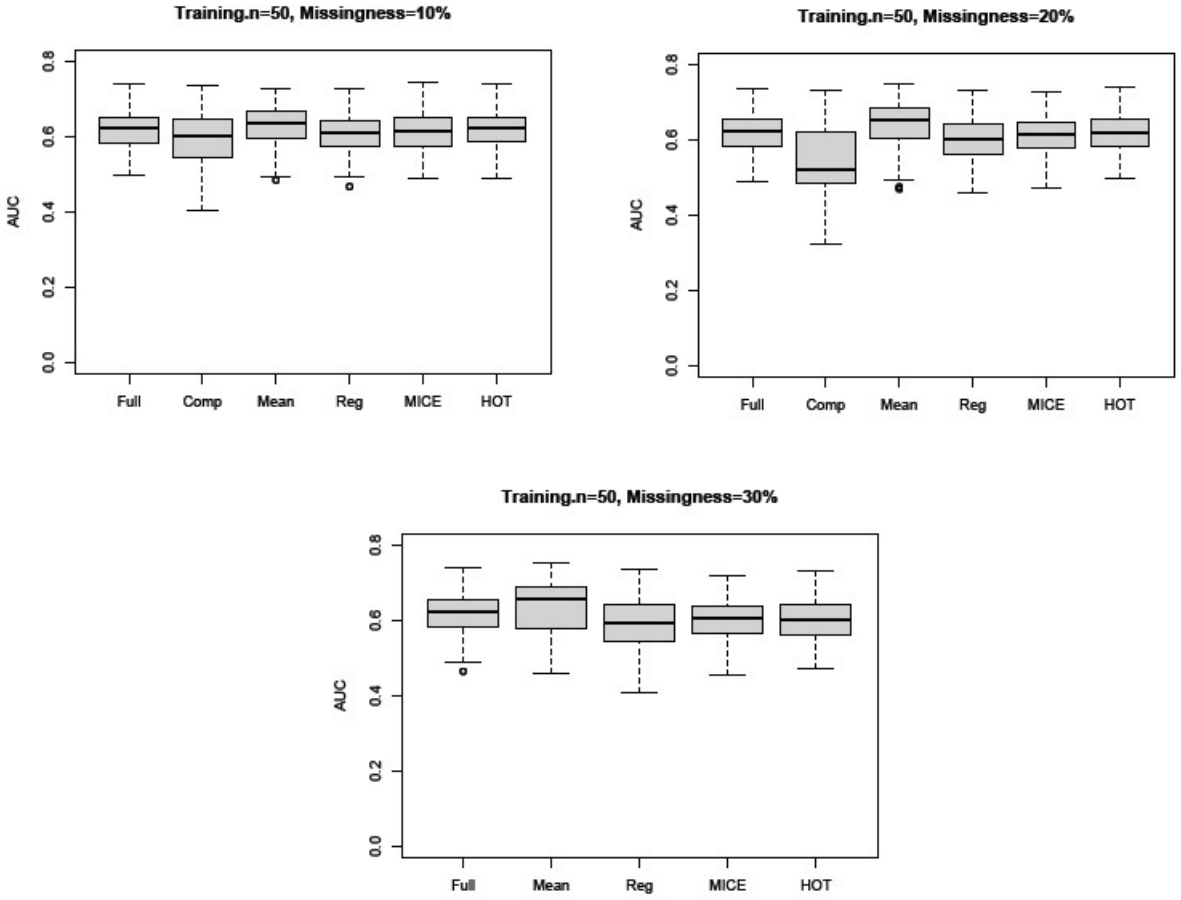

Figure 4.15: Boxplots of the AUC for five imputation methods and for the full data with $N = 50$, POM $= 0.1, 0.2, 0.3$, when the data is MAR.

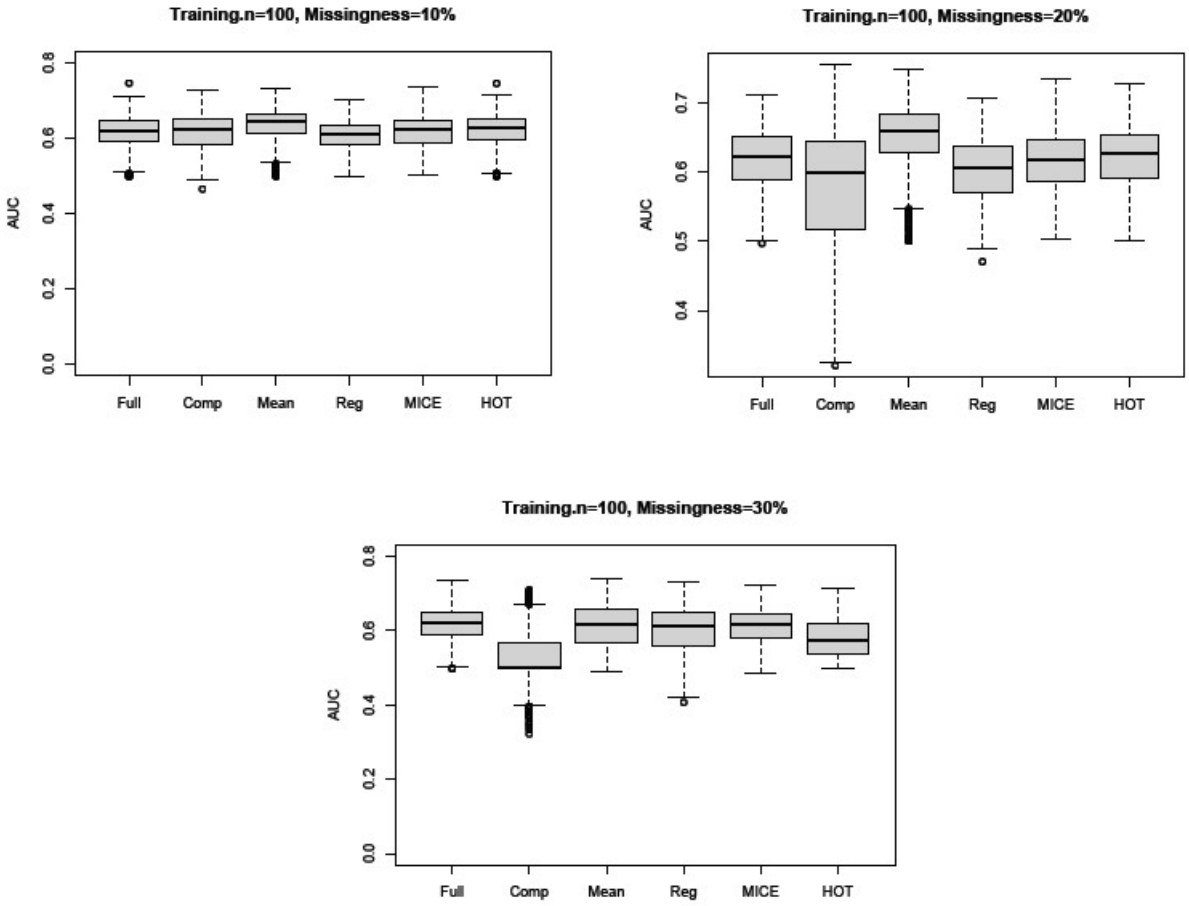

Figure 4.16: Boxplots of the AUC for five imputation methods and for the full data with $N = 100$, POM $= 0.1, 0.2, 0.3$, when the data is MAR.

In Figure 4.16, the AUCs of all of the considered imputation methods were around 0.6, which is close to the full data prediction with POM = 0.1. For POM = 0.2 and POM = 0.3, the AUC of the complete-data-only analysis decreased to approximately 0.5.

4.6.2.3 *A Goodness-of-Fit statistic.* Figures 4.17-4.19 show the p-values of the Hosmer-Lemeshow test. Similar to the result in Figure 4.8-4.10, all of the p-values were above 0.6 for $N = 30$ and above 0.5 for $N = 50$ and $N = 100$ with all three levels of POM. This result indicated that all of our methods fit well.

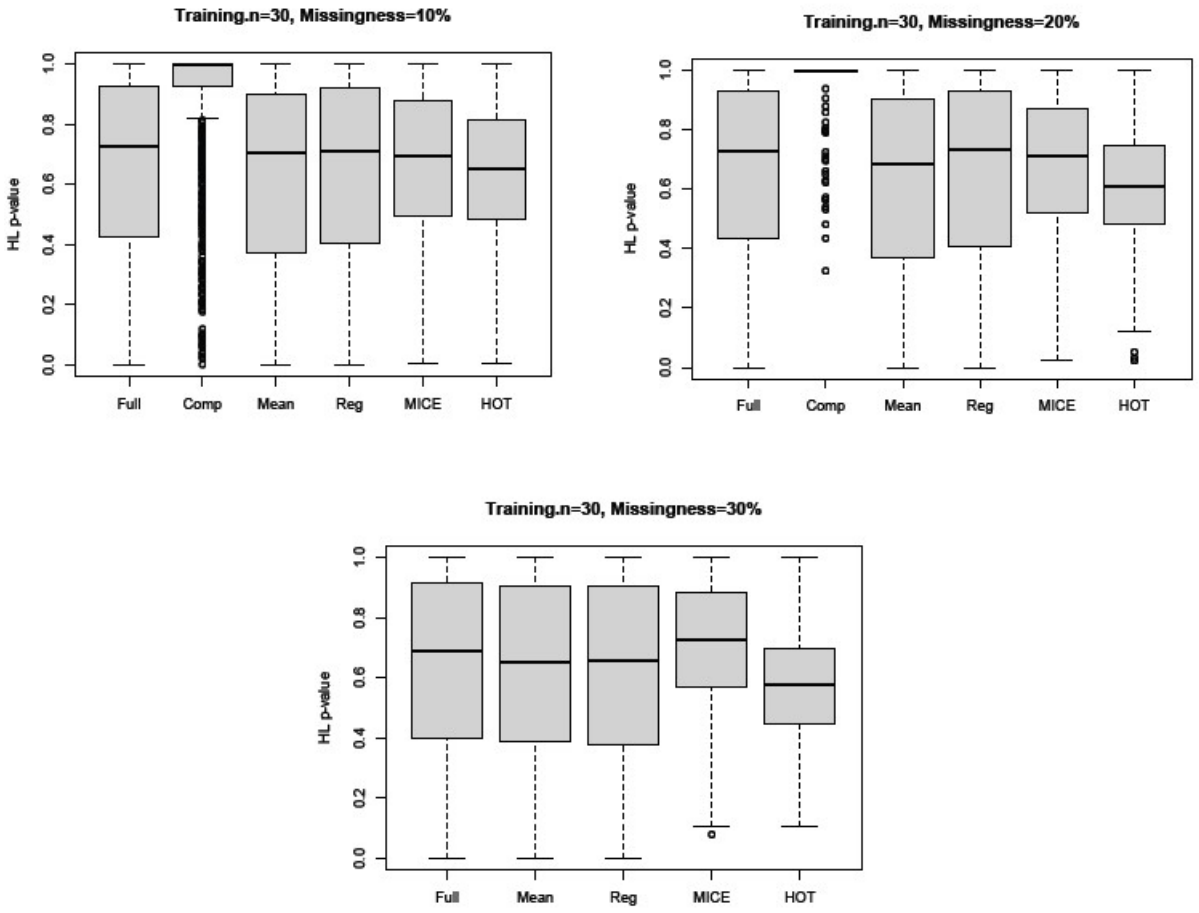

Figure 4.17: Boxplots of the Hosmer-Lemeshow p-value for five imputation methods and for the full data with $N = 30,\ \text{POM} = 0.1, 0.2, 0.3$, when the data is MAR.

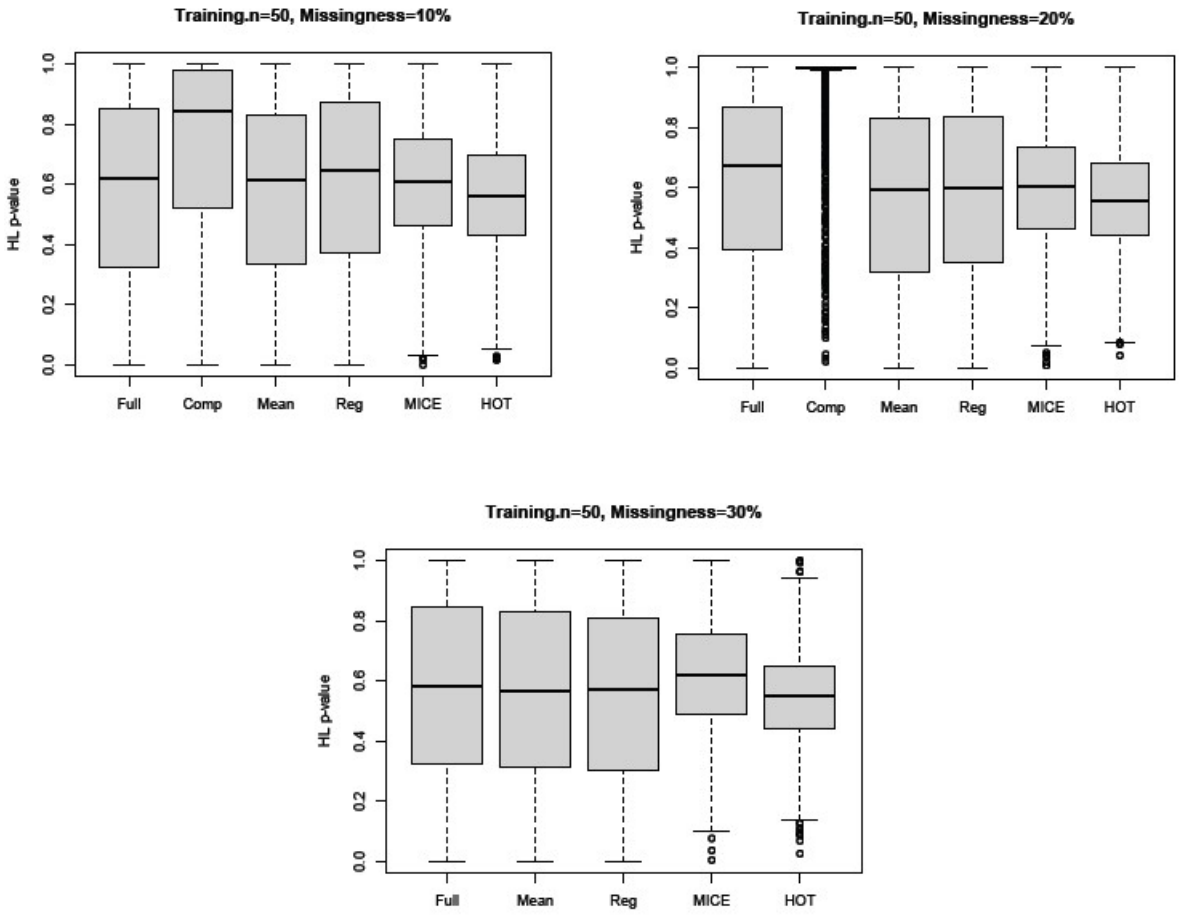

Figure 4.18: Boxplots of the Hosmer-Lemeshow p-value for five imputation methods and for the full data with $N = 50,\ \text{POM} = 0.1, 0.2, 0.3$, when the data is MAR.

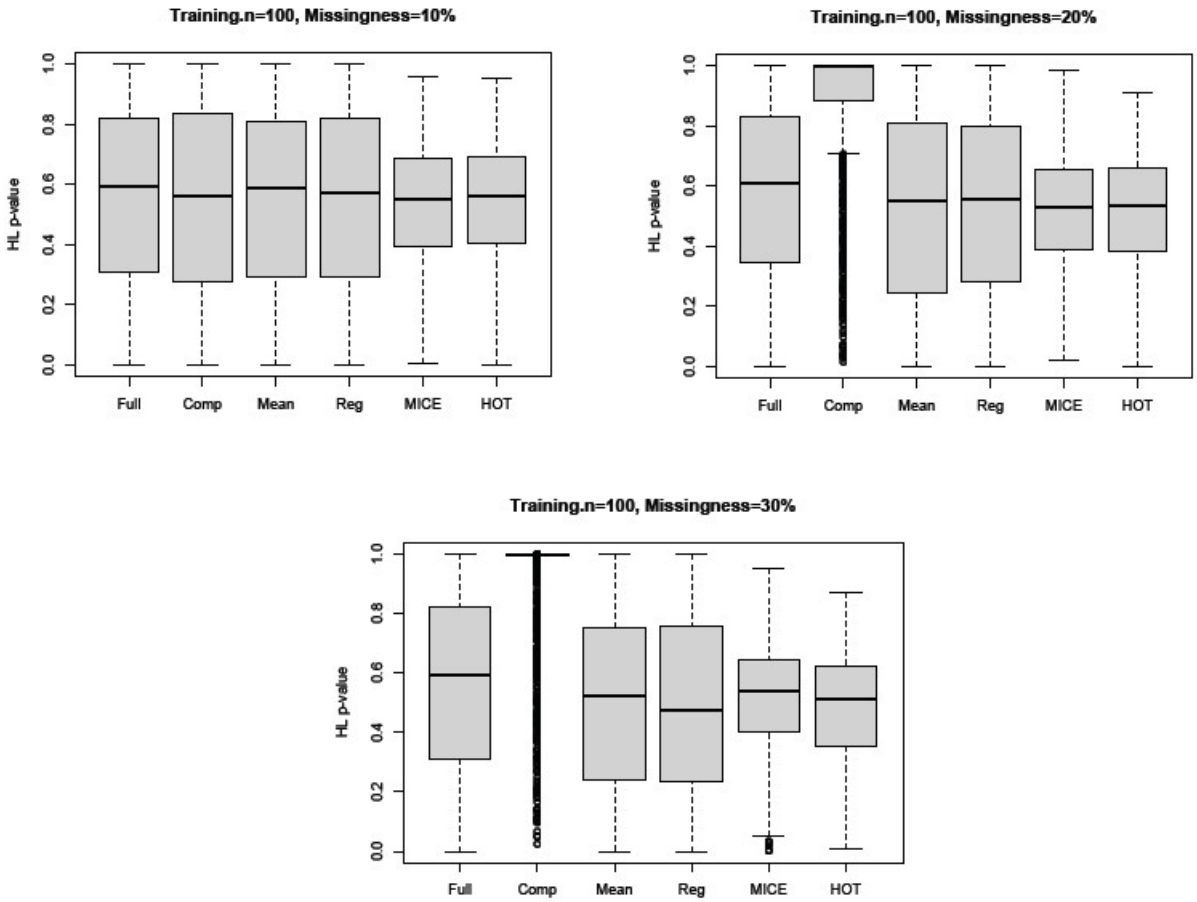

Figure 4.19: Boxplots of the Hosmer-Lemeshow p-value for five imputation methods and for the full data with $N = 100$, POM $= 0.1, 0.2, 0.3$, when the data is MAR.

4.7 Discussion

Missing data often occurs in research with real data. We can either keep the data as it is or replace the missing spots with imputed values. Hence, how we handle missing data becomes the first question that we must answer when performing a statistical analysis. In the presence of missing data, many imputation methods are available for use. The relevant question becomes which imputation method should one choose.

This study compared five imputation methods with various combinations of sample size and percentage of missingness. The considered methods are the following:

- Complete-data-only analysis
- Single-mean imputation
- Conditional-mean imputation
- Multiple imputation by chained equations
- Hotdeck imputation

For the small-sample-size scenarios, the complete-data-only analysis performed poorly compared to all other methods, reflected in terms of a higher MAE. This result is not a surprise because discarding data means discarding information. A common belief among researchers is that multiple imputation is preferable to single-value imputation because multiple imputation has the advantage over single-value imputation of considering the uncertainty of the imputation model. However, in this simulation study, single-mean imputation outperformed other considered single-value imputation methods and MICE.

As the POM increased for a fixed sample size, the MAEs for all considered methods became similar to one another when the POM was small and varied when POM increased. Hotdeck imputation performed the best for POM = 0.3. With a

large sample size and a small POM, the performance of all considered imputation methods we studied was close to one another, and the MAE of complete-data-only analysis increased when the POM increased.

The single-mean imputation and hotdeck imputation methods both performed well with higher levels of POM while the computation of the former is much simpler and easier than that of the latter. Thus, the single-mean imputation might be preferred in practical application since hotdeck imputation is more time-consuming and does not yield a significantly better performance.

APPENDIX

APPENDIX A

Derivation of The Complete Conditional Distributions for the Model Parameters

We use the bivariate normal model for the response (X) and toxicity (Y) random variables. In the two-dimensional nonsingular case, the pdf of the vector $(x_{ij}, y_{ij})'$ is

$$f(x_{ij}, y_{ij}) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_{ij}-\mu_{x_i})^2}{\sigma_x^2} + \frac{(y_{ij}-\mu_{y_i})^2}{\sigma_y^2} - \frac{2\rho(x_{ij}-\mu_{x_i})(y_{ij}-\mu_{y_i})}{\sigma_x\sigma_y}\right]\right),$$

where ρ is the correlation between X and Y, $\sigma_x > 0$, and $\sigma_y > 0$ and

$$\boldsymbol{\mu} \equiv \begin{pmatrix} \mu_{x_i} \\ \mu_{y_i} \end{pmatrix} \text{ with } \boldsymbol{\Sigma} \equiv \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

The likelihood function is

$$f(\boldsymbol{x}, \boldsymbol{y}) = (2\pi)^{-N}|\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\left\{\frac{1}{2}\sum_{i=0}^{t}\sum_{j=1}^{n_i}\left[(x_{ij}, y_{ij})' - (\mu_{x_i}, \mu_{y_i})'\right]'\right.$$
$$\left.\times\, \boldsymbol{\Sigma}^{-1}\left[(x_{ij}, y_{ij})' - (\mu_{x_i}, \mu_{y_i})'\right]\right\},$$

where

$x_{ij} \equiv$ toxicity of the j^{th} individual at the i^{th} dosage level;

$y_{ij} \equiv$ response of the j^{th} individual at the i^{th} dosage level;

$n_i \equiv$ number of patients who received the i^{th} treatment;

$t \equiv$ total number of treatments;

$N \equiv$ total number of patients tested;

$\mu_{x_i} \equiv$ the mean toxicity for the i^{th} treatment;

μ_{x_0} represents the mean toxicity of the placebo

$\mu_{y_i} \equiv$ the mean response for the i^{th} treatment;

μ_{y_0} represents the mean response of the placebo

$\sigma_x^2 \equiv$ the variance of individuals about the mean toxicity at each dosage level;

$\sigma_y^2 \equiv$ the variance of individuals about the mean response at each dosage level.

The prior distributions are

$$
\begin{aligned}
(\mu_{x_0}, \mu_{y_0})' &\sim N\left[\mathbf{0}, \boldsymbol{B}\right], \\
(\mu_{x_i}, \mu_{y_i})' &\sim N\left[(\mu_{x_{i-1}}, \mu_{y_{i-1}})', d_i \boldsymbol{\Sigma}_\tau\right], \\
\boldsymbol{\Sigma} &\sim IW(\boldsymbol{R}, \nu),
\end{aligned}
$$

and

$$
\boldsymbol{\Sigma_\tau} \sim IW(\boldsymbol{R}_\tau, \nu_\tau).
$$

where ν, $\nu_\tau > 0$. To develop a computational approach to estimate the model parameters, we first solve for the complete conditional distribution of each unknown parameter. Hence,

$$
\begin{aligned}
p(\theta|\boldsymbol{x}, \boldsymbol{y}) &\propto p(\theta) p(\boldsymbol{x}, \boldsymbol{y}|\theta) \\
&= (2\pi)^{-N} \boldsymbol{\Sigma}^{-\frac{N}{n}} \exp\left\{ \frac{1}{2} \sum_{i=0}^{t} \sum_{j=1}^{n_i} \left[(x_{ij}, y_{ij})' - (\mu_{x_i}, \mu_{y_i})'\right]' \boldsymbol{\Sigma}^{-1} \right. \\
&\quad \left. \times \left[(x_{ij}, y_{ij})' - (\mu_{x_i}, \mu_{y_i})'\right] \right\} \\
&\quad \times N\left[(\mu_{x_0}, \mu_{y_0})'|0, \boldsymbol{B}\right] \times N\left[(\mu_{x_i}, \mu_{y_i})'|(\mu_{x_{i-1}}, \mu_{y_{i-1}})', d_i \boldsymbol{\Sigma}_\tau\right] \\
&\quad \times N\left[(\mu_{x_{i-1}}, \mu_{u_{i-1}})'|(\mu_{x_{i-2}}, \mu_{y_{i-2}})', d_{i-1} \boldsymbol{\Sigma}_\tau\right] \\
&\quad \times \cdots\cdots \\
&\quad \times N\left[(\mu_{x_t}, \mu_{y_t})'|(\mu_{x_{t-1}}, \mu_{y_{t-1}})', d_{i-1} \boldsymbol{\Sigma}_\tau\right] \\
&\quad \times IW(\boldsymbol{\Sigma}|\boldsymbol{R}, \nu) \\
&\quad \times IW(\boldsymbol{\Sigma}_\tau|\boldsymbol{R}_\tau, \nu_\tau)
\end{aligned}
$$

We first derive the conditional distribution for $(\mu_{x_0}, \mu_{y_0})'$. We have

$$
\begin{aligned}
p\left[(\mu_{x_0}, \mu_{y_0})' | \boldsymbol{x}, \boldsymbol{y}\right] \propto \exp\Big\{ & -\frac{1}{2} \sum_{1}^{n_i} \left[(x_{0j}, y_{0j})' - (\mu_{x_0}, \mu_{y_0})'\right]' \boldsymbol{\Sigma}^{-1} \\
& \times \left[(x_0, y_0)' - (\mu_{x_0}, \mu_{y_0})'\right] \Big\} \\
& \times \exp\Big\{ -\frac{1}{2} \left[(\mu_{x_0}, \mu_{y_0})' - (0,0)'\right]' \boldsymbol{B}^{-1} \left[(\mu_{x_0}, \mu_{y_0})' - (0,0)'\right] \Big\} \\
& \times \exp\Big\{ -\frac{1}{2} \left[(\mu_{x_1}, \mu_{y_1})' - (\mu_{x_0}, \mu_{y_0})'\right]' (d_i \boldsymbol{\Sigma}_\tau)^{-1} \\
& \times \left[(\mu_{x_1}, \mu_{y_1})' - (\mu_{x_0}, \mu_{y_0})'\right] \Big\} \\
= \exp\Big\{ & \sum_{j=1}^{n_0} \left[(x_{0j}, y_{0j})' - (\mu_{x_0}, \mu_{y_0})'\right]' \boldsymbol{\Sigma}^{-1} \\
& \times \left[(x_{0j}, y_{0j})' - (\mu_{x_0}, \mu_{y_0})'\right] \\
& + (\mu_{x_0}, \mu_{y_0})'^T \boldsymbol{B}^{-1} (\mu_{x_0}, \mu_{y_0})' \\
& + \left[(\mu_{x_1}, \mu_{y_1})' - (\mu_{x_0}, \mu_{y_0})'\right]' (d_i \boldsymbol{\Sigma}_\tau)^{-1} \left[(\mu_{x_1}, \mu_{y_1})' - (\mu_{x_0}, \mu_{y_0})'\right] \Big\} \\
= \exp\Big\{ & \sum_{j=1}^{n_0} \big[(x_{0j}, y_{0j}) \boldsymbol{\Sigma}^{-1} (x_{0j}, y_{0j})' - (x_{0j}, y_{0j}) \boldsymbol{\Sigma}^{-1} (\mu_{x_0}, \mu_{y_0})' \\
& - (\mu_{x_0}, \mu_{y_0}) \boldsymbol{\Sigma}^{-1} (x_{0j}, y_{0j})' + (\mu_{x_0}, \mu_{y_0}) \boldsymbol{\Sigma}^{-1} (\mu_{x_0}, \mu_{y_0})' \\
& + (\mu_{x_0}, \mu_{y_0}) \boldsymbol{B}^{-1} (\mu_{x_0}, \mu_{y_0})' \\
& + (\mu_{x_1}, \mu_{y_1}) (d_i \boldsymbol{\Sigma}_\tau)^{-1} (\mu_{x_1}, \mu_{y_1}) - (\mu_{x_0}, \mu_{y_0}) (d_i \boldsymbol{\Sigma}_\tau)^{-1} (\mu_{x_1}, \mu_{y_1})' \\
& - (\mu_{x_1}, \mu_{y_1}) (d_i \boldsymbol{\Sigma}_\tau)^{-1} (\mu_{x_0}, \mu_{y_0})' + (\mu_{x_0}, \mu_{y_0}) (d_i \boldsymbol{\Sigma}_\tau)^{-1} (\mu_{x_0}, \mu_{y_0})' \big] \Big\} \\
\propto \exp\Big\{ & (\mu_{x_0}, \mu_{y_0}) \left[n_0 \boldsymbol{\Sigma}^{-1} + \boldsymbol{B}^{-1} + (d_i \boldsymbol{\Sigma}_\tau)^{-1}\right] (\mu_{x_0}, \mu_{y_0})' \\
& - \left[n_0 (\bar{x}_{0\cdot}, \bar{y}_{0\cdot}) \boldsymbol{\Sigma}^{-1} + (\mu_{x_1}, \mu_{y_1}) (d_i \boldsymbol{\Sigma}_\tau)^{-1}\right] (\mu_{x_0}, \mu_{y_0})' \\
& - (\mu_{x_0}, \mu_{y_0}) \left[n_0 \boldsymbol{\Sigma}^{-1} (\bar{x}_{0\cdot}, \bar{y}_{0\cdot})' + (d_i \boldsymbol{\Sigma}_\tau)^{-1} (\mu_{x_1}, \mu_{y_1})'\right] \Big\}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
(\mu_{x_0}, \mu_{y_0})' \sim N\Big\{ & [n_0 \boldsymbol{\Sigma}^{-1} + \boldsymbol{B}^{-1} + (d_1 \boldsymbol{\Sigma}_\tau)^{-1}]^{-1} [n_0 (\bar{X}_0)' \boldsymbol{\Sigma}^{-1} + \mu_1' (d_1 \boldsymbol{\Sigma}_\tau)^{-1}], \\
& [n_0 \boldsymbol{\Sigma}^{-1} + \boldsymbol{B}^{-1} + (d_1 \boldsymbol{\Sigma}_\tau)^{-1}]^{-1} \Big\}.
\end{aligned}
$$

Next, we derive the conditional distribution for $(\mu_{x_t}, \mu_{y_t})'$. We have

$$
\begin{aligned}
p\left[(\mu_{x_t}, \mu_{y_t})'|X, Y\right] \propto \exp\Big\{ & -\frac{1}{2}\sum_{1}^{n_i}\left[(x_{tj}, y_{tj})' - (\mu_{x_t}, \mu_{y_t})'\right]' \boldsymbol{\Sigma}^{-1} \\
& \times \left[(x_{tj}, y_{tj})' - (\mu_{x_t}, \mu_{y_t})'\right] \Big\} \\
& \times \exp\Big\{ -\frac{1}{2}\left[(\mu_{x_t}, \mu_{y_t})' - (\mu_{x_{t-1}}, \mu_{y_{t-1}})'\right]' (d_t\boldsymbol{\Sigma}_\tau)^{-1} \\
& \times \left[(\mu_{x_t}, \mu_{y_t})' - (\mu_{x_{t-1}}, \mu_{y_{t-1}})'\right] \Big\} \\
= \exp\Big\{ & -\frac{1}{2}\sum_{j=1}^{n_j}\left[(x_{tj}, y_{tj})' - (\mu_{x_t}, \mu_{y_t})'\right]' \boldsymbol{\Sigma}^{-1}\left[(x_{tj}, y_{tj})'\right. \\
& \left.-(\mu_{x_t}, \mu_{y_t})'\right] - \left[(\mu_{x_t}, \mu_{y_t})' - (\mu_{x_{t-1}}, \mu_{y_{t-1}})'\right]' d_t\boldsymbol{\Sigma}_\tau \\
& \times \left[(\mu_{x_t}, \mu_{y_t})' - (\mu_{x_{t-1}}, \mu_{y_{t-1}})'\right] \Big\} \\
\propto & (\mu_{x_t}, \mu_{y_t})\left[(d_t\boldsymbol{\Sigma}_\tau)^{-1} + \boldsymbol{\Sigma}^{-1}\right](\mu_{x_t}, \mu_{y_t})' - (\mu_{x_t}, \mu_{y_t}) \\
& \times \left[\boldsymbol{\Sigma}^{-1} n_t(\bar{x}_t, \bar{y}_t)' + (d_t\boldsymbol{\Sigma}_\tau)^{-1}(\mu_{x_{t-1}}, \mu_{y_{t-1}})'\right].
\end{aligned}
$$

Hence, the conditional distribution of $(\mu_{x_t}, \mu_{y_t})'$ is

$$
\begin{aligned}
(\mu_{x_t}, \mu_{y_t})' \sim N\Big\{ & [(d_t\boldsymbol{\Sigma}_\tau)^{-1} + n_t\boldsymbol{\Sigma}^{-1}]^{-1}[\boldsymbol{\Sigma}^{-1} n_t(\bar{X}_t) + (d_t\boldsymbol{\Sigma}_\tau)^{-1}\mu_{t-1}], \\
& [(d_t\boldsymbol{\Sigma}_\tau)^{-1} + n_t\boldsymbol{\Sigma}^{-1}]^{-1}\Big\}.
\end{aligned}
$$

The derivation of the conditional distribution of $(\mu_{x_i}, \mu_{y_i})'$ is very similar to that of $(\mu_{x_t}, \mu_{y_t})'$ and results in

$$
\begin{aligned}
(\mu_{x_t}, \mu_{y_t})' \sim N\Big\{ & [n_i\boldsymbol{\Sigma}^{-1} + (d_i\boldsymbol{\Sigma}_\tau)^{-1} + (d_{i+1}\boldsymbol{\Sigma}_\tau)^{-1}]^{-1} \\
& \times \left[\mu'_{i-1}(d_i\boldsymbol{\Sigma}_\tau)^{-1} + n_i(\bar{X}_i)'\boldsymbol{\Sigma}^{-1} + \mu'_{i+1}(d_{i+1}\boldsymbol{\Sigma}_\tau)^{-\mathbf{1}}\right], \\
& [n_i\boldsymbol{\Sigma}^{-1} + (d_i\boldsymbol{\Sigma}_\tau)^{-1} + (d_{i+1}\boldsymbol{\Sigma}_\tau)^{-1}]^{-1}\Big\}.
\end{aligned}
$$

Because the Inverse-Wishart is a conjugate prior for the normal likelihood, the conditional distributions of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_\tau$ are

$$
\boldsymbol{\Sigma} \sim IW\Big\{N + \nu, \boldsymbol{R} + \textstyle\sum_{i=0}^{t}\sum_{j=1}^{n_i}\left[(x_{ij}, y_{ij})' - (\mu_{x_i}, \mu_{y_i})'\right]\left[(x_{ij}, y_{ij})' - (\mu_{x_i}, \mu_{y_i})'\right]'\Big\}
$$

and

$$\boldsymbol{\Sigma}_\tau \sim IW\Big\{\tfrac{t}{2}+\nu_\tau, \boldsymbol{R}_\tau + \textstyle\sum_{i=1}^{t}\left[(\mu_{x_i},\mu_{y_i})' - (\mu_{x_{i-1}},\mu_{y_{i-1}})'\right]\left[(\mu_{x_i},\mu_{y_i})' - (\mu_{x_{i-1}},\mu_{y_{i-1}})'\right]'\Big\}.$$

Thus, the complete conditional distributions of the parameters are

$$\begin{aligned}(\mu_{x_0},\mu_{y_0})' \sim N\Big\{&[n_0\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}^{-1} + (d_1\boldsymbol{\Sigma}_\tau)^{-1}]^{-1}[n_0(\bar{X}_0)'\boldsymbol{\Sigma}^{-1} + \boldsymbol{\mu}_1'(d_1\boldsymbol{\Sigma}_\tau)^{-1}],\\ &[n_0\boldsymbol{\Sigma}^{-1} + \boldsymbol{B}^{-1} + (d_1\boldsymbol{\Sigma}_\tau)^{-1}]^{-1}\Big\},\end{aligned}$$

$$\begin{aligned}(\mu_{x_t},\mu_{y_t})' \sim N\Big\{&[(d_t\boldsymbol{\Sigma}_\tau)^{-1} + n_t\boldsymbol{\Sigma}^{-1}]^{-1}[\boldsymbol{\Sigma}^{-1}n_t(\bar{X}_t) + (d_t\boldsymbol{\Sigma}_\tau)^{-1}\boldsymbol{\mu}_{t-1}],\\ &[(d_t\boldsymbol{\Sigma}_\tau)^{-1} + n_t\boldsymbol{\Sigma}^{-1}]^{-1}\Big\},\end{aligned}$$

$$\begin{aligned}(\mu_{x_i},\mu_{y_i})' \sim N\Big\{&[n_i\boldsymbol{\Sigma}^{-1} + (d_i\boldsymbol{\Sigma}_\tau)^{-1} + (d_{i+1}\boldsymbol{\Sigma}_\tau)^{-1}]^{-1}\left[\boldsymbol{\mu}_{i-1}'(d_i\boldsymbol{\Sigma}_\tau)^{-1} + n_i(\bar{X}_i)'\boldsymbol{\Sigma}^{-1}\right.\\ &\left.+\boldsymbol{\mu}_{i+1}'(d_{i+1}\boldsymbol{\Sigma}_\tau)^{-1}\right], [n_i\boldsymbol{\Sigma}^{-1} + (d_i\boldsymbol{\Sigma}_\tau)^{-1} + (d_{i+1}\boldsymbol{\Sigma}_\tau)^{-1}]^{-1}\Big\},\end{aligned}$$

$$\boldsymbol{\Sigma} \sim IW\Big\{N+\nu, \boldsymbol{R} + \sum_{i=0}^{t}\sum_{j=1}^{n_i}\left[(x_{ij},y_{ij})' - (\mu_{x_i},\mu_{y_i})'\right]\left[(x_{ij},y_{ij})' - (\mu_{x_i},\mu_{y_i})'\right]'\Big\},$$

and

$$\boldsymbol{\Sigma}_\tau \sim IW\Big\{\tfrac{t}{2}+\nu_\tau, \boldsymbol{R}_\tau + \textstyle\sum_{i=1}^{t}\left[(\mu_{x_i},\mu_{y_i})' - (\mu_{x_{i-1}},\mu_{y_{i-1}})'\right]\left[(\mu_{x_i},\mu_{y_i})' - (\mu_{x_{i-1}},\mu_{y_{i-1}})'\right]'\Big\}.$$

BIBLIOGRAPHY

Agresti, A. (2012), *Categorical Data Analysis*, Wiley, 3rd ed.

Ambler, G., Omar, R., and Royston, P. (2007), "A Comparison of Imputation Techniques for Handling Missing Predictor Values in a Risk Model with a Binary Outcome," *Statistical Methods in Medical Research*, 16, 277–298.

Carlin, B. and Louis, T. (2008), *Bayesian Methods for Data Analysis*, CRC Press, 3rd ed.

FDA (2010), *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics*, Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research.

Gelman, A. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 5th ed.

Geman, S. and Geman, D. (1984), "An Adaptive Bayesian Approach to Continuous Dose-Response Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Goodman, S., Zahurak, M., and Piantadosi, S. (1995), "Some Practical Improvements in the Continual Reassessment Method for Phase I Studies," *Statistics in Medicine*, 14, 1149–1161.

Gooley, T., Martin, P., Fisher, L., 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.

Hastings, W. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Hoering, A., Leblanc, M., and Crowley, J. (2011), "Seamless Phase I/II Trial Design for Assessing Toxicity and Efficacy for Targeted Agents," *Clinical Cancer Research*, 17, 640–646.

Kairalla, J., Coffey, C., Thomann, M., and Muller, K. (2012), "Adaptive Trial Designs: A Review of Barriers and Opportunities," *Trials*, 13, 145–153.

Kaw, A., Kalu, E., and Nguyen, D. (2011), *Numerical Methods with Applications*, University of South Florida, 1st ed.

Koopmeiners, J. and Modiano, J. (2013), "A Bayesian Adaptive Phase I-II Clinical Trial for Evaluating Efficacy and Toxicity with Delayed Outcomes," *Clinical Trials*, 0, 1–11.

Korn, E., Midthune, D., Chen, T., Rubinstein, L., Christian, M., and Simon, R. (1994), "A Comparison of Two Phase I Trial Designs," *Statistics in Medicine*, 13, 1799–1806.

Leininger, T. (2010), "An Adaptive Bayesian Approach to Continuous Dose-Response Modeling," Master's thesis, Brigham Young University.

Lewis, R., Viele, K., Broglio, K., Berry, S., and Jones, A. (2013), "An Adaptive, Phase II, Dose-finding Clinical Trial Design to Evaluate L-Carnitine in the Treatment of Septic Shock Based on Efficacy and Predictive Probability of Subsequent Phase III Success," *Critical Care Medicine*, 41, 1674–1678.

Mtropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, 21, 1087–1091.

Steyerberg, E., Vickers, A., Cook, N., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M., and Kattan, M. (2010), "Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures," *Epidemiology*, 21, 128–138.

Tanner, M. (1998), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, New York: Springer-Verlag, 2nd ed.

Thall, P. and Cook, J. (2004), "Dose-Finding Based on Efficacy-Toxicity Trade-Offs," *Biometrics*, 60, 684–693.

Thall, P., Cook, J., and Estey, E. (2007), "Adaptive Dose Selection Using Efficacy-Toxicity Trade-Offs: Illustrations and Practical Considerations," *Journal of Biopharmaceutical Statistics*, 16:5, 623–638.

Thall, P., Estey, E., and Sung, H. (1999), "A New Statistical Method for Dose-Finding Based on Efficacy and Toxicity in Early Phase Clinical Trials," *Investigational New Drugs*, 17, 155–167.

Thall, P. and Russell, K. (1998), "A Strategy for Dose-Finding and Satety Monitoring Based on Efficacy and Adverse Outcomes in Phase I/II Clinical Trials," *Biometrics*, 54, 251–264.

Walter, S. (2005), "The Partial Area under the Summary ROC Curve," *Statistics In Medicine*, 24, 2025–2040.

Zhang, W., Sargent, D., and Mandrekar, S. (2006), "An Adaptive Dose-Finding Design Incorporating both Toxicity and Efficacy," *Statistics in Medicine*, 25, 2365–2383.