A TWO-STAGE COVARIATE-ADJUSTED RESPONSE-ADAPTIVE ENRICHMENT DESIGN

by

Li Yang

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Statistical Science

Committee:

_____	Dr. William F. Rosenberger, Dissertation Director
_____	Dr. Guoqing Diao, Dissertation Co-Director
_____	Dr. Daniel B. Carr, Committee Member
_____	Dr. Clifton D. Sutton, Committee Member
_____	Dr. Estelle Russek-Cohen, Committee Member
_____	Dr. William F. Rosenberger, Department Chair
_____	Dr. Kenneth Ball, Dean, The Volgenau School of Engineering

Date: _____ Summer Semester 2019
George Mason University
Fairfax, VA

A Two-Stage Covariate-Adjusted Response-Adaptive Enrichment Design

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Li Yang
Master of Science
George Mason University, 2008
Bachelor of Medicine
Central South University, 2000

Director: Dr. William F. Rosenberger, Professor
Co-Director: Dr. Guoqing Diao, Professor
Department of Statistics

Summer Semester 2019
George Mason University
Fairfax, VA

Dedication

To my husband Frank and my three children Daniel, Albert, and Ellen Zeng.

Acknowledgments

First and foremost, I would like to express my deepest appreciation to my advisors Dr. William F. Rosenberger and Dr. Guoqing Diao, for their guidance, support, encouragement, and patience throughout the years. I greatly appreciate their decisions to become my mentors, their step-by-step guidance throughout my research, and their flexibility and understanding to my work-life schedule.

I am also sincerely grateful to my committee members, Dr. Daniel B. Carr, Dr. Clifton D. Sutton, and Dr. Estelle Russek-Cohen, for their valuable comments, encouragement and support. In particular, I am very appreciate Dr. Russek-Cohen's consideration to meet me near my office and her guidance in enrichment designs. I am grateful that Dr. Carr took the time out of his busy schedule to review data visualization materials and share his experience with me. I thank Dr. Sutton for his timely responses, thoughtful comments, and flexible schedule.

I would also like to extend my sincere gratitude to my supervisor at NIH Dr. Gwenyth R. Wallen for her support over the years. The completion of my doctoral study would not have been possible without her support. I also thank my colleague Dr. Alyssa T. Brooks for proofreading and correcting my dissertation. I would like to thank Ms. Elizabeth Quigley and Ms. Verronica Mitchell for helping me schedule numerous meetings and booking conference rooms.

Very special thanks to my husband Frank for encouraging me to pursue the academic degree and giving me unconditional support over the years. I also would like to thank my children Daniel, Albert, and Ellen for cheering me up, taking care of themselves, and helping household chores.

Table of Contents

Page

List of Tables vii
List of Figures ix
Abstract x
1 Introduction and Literature Review 1
1.1 General Definition of CARA 1
1.2 CARA in the Literature 2
1.2.1 Treatment Effect Mapping 3
1.2.2 Target Allocation Approach 4
1.2.3 Weighted Optimality Approach 6
1.2.4 Bayesian Adaptive Randomization Methods 11
1.2.5 Non-parametric CARA Procedures 11
1.3 Properties of the GLM Approach to CARA 12
1.3.1 General CARA Procedure Framework and Asymptotic Properties 12
1.3.2 Generalized Linear Model 16
1.4 General Definition of Adaptive Enrichment Design 19
1.5 Adaptive Enrichment Designs in the Literature 21
1.5.1 Strata-Based Adaptive Enrichment Design 21
1.5.2 Model-Based Adaptive Enrichment Design 23
1.6 Simon's Adaptive Enrichment Design 26
1.6.1 Adaptive Enrichment Design for Two Group Binary Outcome 26
1.6.2 Adaptive Threshold Enrichment Design 28
1.6.3 Group Sequential Analysis 28
1.7 Randomization Tests 30
1.8 Contribution and Outline of the Thesis 32
2 Comparison of Different CARA Procedures 33
2.1 Binary outcomes in logistic regression models 33
2.2 Continuous outcomes in linear regression models 41

2.3 Conclusion 45
3 Monte Carlo Tests of Interaction Effect 48
3.1 Binary responses 50
3.2 Continuous responses 52
3.3 Conclusion 53
4 A two-stage Enrichment Design using Monte Carlo Tests 56
4.1 Binary responses 58
4.2 Continuous responses 60
4.3 Redesigning an existing trial 65
4.4 Conclusion 71
5 Conclusions and Future Work 72
Bibliography 75

List of Tables

Table		Page
2.1	Type I error rates from different randomization designs for two equally distributed covariate strata and binary outcomes	38
2.2	Power, allocation ratios, and overall success rates from different randomization designs for two equally distributed covariate strata and binary outcomes	40
2.3	Power, allocation ratios, and overall success rates from different randomization designs for two unequally (2 : 8) distributed covariate strata and binary outcomes	42
2.4	Type I error rates from different randomization designs for four equally distributed covariate strata and continuous outcomes with 10% outliers	45
2.5	Power and allocation ratios from different randomization designs for four equally distributed covariate strata and continuous outcomes	46
2.6	Power and allocation ratios from different randomization designs for four equally distributed covariate strata and continuous outcomes with 10% outliers	46
4.1	Power and overall success rates from different designs for two equally distributed covariate strata and binary outcomes	60
4.2	Power and overall success rates from different designs for two unequally (2 : 8) distributed covariate strata and binary outcomes	61
4.3	Power and overall success rates from different designs for four equally distributed covariate strata and binary outcomes	62
4.4	Type I error rates and powers from different designs for two equally distributed covariate strata and continuous outcomes	65
4.5	Type I error rates and powers from different designs for two equally distributed covariate strata and continuous outcomes with 10% outliers	66

4.6 Type I error rates and powers from different designs for four equally distributed covariate strata and continuous outcomes 67
4.7 Type I error rates and powers from different designs for four equally distributed covariate strata and continuous outcomes with 10% outliers 68
4.8 Success rates and p-value from different designs based on NSABP trial 70

List of Figures

Figure		Page
2.1	Power from different CARA procedures, 5000 runs, n = 1000	39
2.2	Power from different DBCD procedures, 5000 runs, n = 400	47
3.1	Two covariate strata, two treatment groups, and binary outcomes . .	51
3.2	Four covariate strata, two treatment groups, and binary outcomes . .	52
3.3	Two covariate strata, two treatment groups, and continuous outcomes	54
3.4	Four covariate strata, two treatment groups, and continuous outcomes	55
4.1	Powers using different types of data for binary outcomes	63
4.2	Powers using different types of data for continuous outcomes	69

Abstract

A TWO-STAGE COVARIATE-ADJUSTED RESPONSE-ADAPTIVE ENRICHMENT DESIGN

Li Yang, PhD

George Mason University, 2019

Dissertation Directors: Dr. William F. Rosenberger and Dr. Guoqing Diao

With the rapid development in genomic and genetic research, precision medicine has gained more attention in modern clinical trials. Molecularly targeted therapies are likely to only work with a subgroup of patients. However, the subgroup often will not be identified until after a large scale clinical trial. Clinical trials are often designed under the assumption of no treatment-by-covariate interaction effect and enroll all comers. This makes many patients go through unnecessary treatment and may decrease the efficiency of the trial.

In this dissertation, we propose a novel two-stage enrichment design which uses covariate-adjusted response-adaptive (CARA) randomization and a Monte Carlo test to evaluate the interaction effect in the interim analysis for binary and continuous outcomes. A pre-defined alpha level is used as the threshold to decide whether a subgroup will be identified and recruited in the second stage. If a below-threshold interaction effect is found, a regression model will be fitted and the stratum with the largest treatment effect will be chosen as the best stratum. The trial will continue to the second stage with patients from the best stratum only. If the p-value from the

interim analysis is above the threshold, the trial continues with all patients. The primary aim is to test the treatment effect between treatment groups. Different CARA procedures are compared in terms of type I error rates, power, and ethical considerations. The CARA procedure that balances better between efficiency and ethics is used in the proposed two-stage enrichment design.

Chapter 1: Introduction and Literature Review

1.1 General Definition of CARA

A clinical trial refers to any research study that prospectively assigns human subjects to one or more interventions to evaluate the effects on health outcomes. Participants in clinical trials are typically randomly assigned to one of those interventions. Randomization promotes comparability with respect to both known and unknown covariates among groups (Rosenberger and Lachin, 2015).

Hu and Rosenberger (2006) describe five classes of randomization procedures: complete randomization, restricted randomization, covariate-adaptive randomization, response-adaptive randomization, and covariate-adjusted response-adaptive (CARA) randomization. Consider a clinical trial with n patients, each of whom is randomly assigned to one of K groups.

A randomization sequence is a matrix $\boldsymbol{T} = (\boldsymbol{T}_1, \cdots, \boldsymbol{T}_n)'$, where $\boldsymbol{T}_i = \boldsymbol{e}_j, j = 1, \cdots, K, i = 1, \cdots, n$ and $\boldsymbol{e}_j$ is a vector with 1 in the j-th position and all other zeros. A response sequence is a matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)'$. Let $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n$ be a set of covariates, let $\mathcal{T}_n = \sigma\{\boldsymbol{T}_1, \cdots, \boldsymbol{T}_n\}$ be the sigma-algebra generated by the first n treatment assignments, let $\mathcal{X}_n = \sigma\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$ be the sigma-algebra generated by the first n responses, and let $\mathcal{Z}_n = \sigma\{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n\}$ be the sigma-algebra generated by the first n covariates vectors. Let $\mathcal{F}_n = \mathcal{T}_n \otimes \mathcal{X}_n \otimes \mathcal{Z}_{n+1}$.

For two treatment group trials, complete randomization is simple coin tossing. The treatment assignments $T_1, \cdots, T_n$ are independent and identically distributed

Bernoulli variables with $Pr(T_j = 1) = 1/2$. When the sample size is small or moderate, Rosenberger and Lachin (2015) showed the imbalance can be significant. Therefore complete randomization is rarely used in practice. Restricted randomization calculates the allocation probability based on previous assignments, $\phi_n = E(T_n|\mathcal{F}_{n-1}) = E(T_n|\mathcal{T}_{n-1})$. Because restricted randomization generates equal or nearly equal numbers of patients in each group, this method is widely used in many clinical trials. Sometimes known covariates play a significant role in the outcome. For example, males and females may have very different responses to the tested treatment. In order to minimize covariate imbalances, covariate-adaptive randomization is used to calculate the allocation probability based on previous assignments and all past and current covariate information, $\phi_n = E(T_n|\mathcal{F}_{n-1}) = E(T_n|\mathcal{T}_{n-1}, \mathcal{Z}_n)$. Response-adaptive randomization changes the allocation probability according to previous treatment assignments and responses, $\phi_n = E(T_n|\mathcal{F}_{n-1}) = E(T_n|\mathcal{T}_{n-1}, \mathcal{X}_{n-1})$. The main goal is to assign more patients into superior treatment and maximize the power of the test of the treatment effect. CARA randomization procedures calculate the allocation probability based on the previous responses, treatment assignments, covariates and the current patient's covariates, $\phi_n = E(T_n|\mathcal{F}_{n-1}) = E(T_n|\mathcal{T}_{n-1}, \mathcal{X}_{n-1}, \mathcal{Z}_n)$. A CARA procedure assigns more patients to the superior treatment group based on patients' characteristics. It takes into consideration patient heterogeneity to achieve both ethical and efficiency goals.

1.2 CARA in the Literature

Rosenberger et al. (2012) review four approaches to CARA randomization. The first approach is a treatment effect mapping. Patients are randomized with probabilities

proportional to the current estimate of the treatment difference, adjusting for covariates. The second approach is based on specifying a target randomization function to calculate the desired proportions of patients for different treatment groups and covariate values. The third approach is a weighted optimality approach. Under this approach, treatment allocation probabilities are sequentially calculated by maximizing some utility function that combines ethical and inferential criteria. The fourth approach is a Bayesian approach. Randomization probabilities for any incoming patient are based on some criteria that favor the best treatment group and accommodate the posterior distribution of the parameters and covariates of the new patient.

1.2.1 Treatment Effect Mapping

Rosenberger et al. (2001b) first introduced the concept of CARA. They used a treatment effect mapping approach, which map the current treatment effect (treatment difference between two treatment groups) to $p_n \in [0, 1]$. The incoming patient is assigned to the treatment A with this mapped probability p_n. For a two groups logistic regression model with covariate-treatment interaction: $\text{logit}(p_1) = \alpha+\beta T_i+\gamma Z_i'+\delta T_i Z_i'$, the proposed allocation mapping is given by $p_{i+1} = \{1+\exp(-\hat{\beta}_i+\hat{\delta}_i Z_{i+1}')\}^{-1}$. Staggered entry and delayed response mechanism are used in the simulations. the proposed method has similar power as equal allocation and fewer treatment failures.

Bandyopadhyay and Biswas (2001) considered a CARA procedure for two treatment group continuous responses. The proposed allocation rule for testing the treatment effect is:

$$\pi_{i+1}(\theta_i, Z_i) = \Phi\left(\frac{\hat{\mu}_{Ai} - \hat{\mu}_{Bi}}{G}\right),$$

where $\hat{\mu}_{Ai} - \hat{\mu}_{Bi}$ is the estimated treatment effect and G is a tuning parameter. This

design does not incorporate current patient's covariate information in the randomization function. Simulation studies show that more patients are assigned to better treatment group as the treatment difference increases. When there is no treatment difference, patients are allocated to two groups with equal probability. When there is a treatment difference, the smaller the G value, the more the allocation is skewed, but the standard deviation of the allocation proportion is also larger. The power for testing the treatment difference is lower than equal randomization procedures when the sample size is small. The confidence interval for the estimated treatment difference is also wider than for equal allocation randomization procedures.

Zhu et al. (2013) studied a binary treatment-by-covariate interaction effect. They proposed an optimal allocation rule for testing a binary covariate by two group interaction effect. Simulation studies compared their procedure with other traditional methods when covariates are correlated, when one of the covariates changes over time, and when the model is misspecified. The results show that the proposed method has higher power than traditional methods under all scenarios.

1.2.2 Target Allocation Approach

Rosenberger et al. (2001a) derived the optimal allocation between two treatments for binary response trials. The proposed optimal allocation rule is given by

$$\pi_{1,i+1} = \frac{\sqrt{p_{1,i}}}{\sqrt{p_{1,i}} + \sqrt{p_{2,i}}},$$

where $\pi_{1,i+1}$ is the randomization probability to the group 1 for the $i+1$th patient, $p_{1,i}$ is the success rate in the group 1 based on all i patients, $p_{2,i}$ is the success rate in the group 2 based on all i patients. This procedure minimizes the expected number

of threatment failures subject to the fixed variance of the test statistic. The proposed design is compared with the randomized play-the-winner rule, Neyman allocation, and equal allocation procedures. There are situations in which the proposed CARA dssign will result in fewer treatment failures without a sacrifice in power.

Chambaz et al. (2014) focused on a group-sequential CARA randomized controlled trial incorporating more flexible techniques to model the response. They choose the marginal treatment effect between two treatment groups as their parameter of interest. The parameters are estimated using targeted minimum loss estimation on top of the least absolute shrinkage and selection operator methodology. The targeted optimal design is Neyman allocation. They show that under mild assumptions, the resulting randomization sequence converges to a limiting design and the minimum loss estimation estimator is consistent and asymptotically normal with an estimable asymptotic variance.

Zhu (2015) discussed the conditional independence and distribution of allocated responses for two groups CARA procedure. They proposed a new CARA procedure allowing common parameters. Suppose for a given covariate $\boldsymbol{Z}$, the response X_k of the treatment $k = 1, \cdots, K$ has a distribution in the exponential family under the generalized linear model

$$f_k(X_k|\boldsymbol{Z}, \boldsymbol{\theta}_k, \boldsymbol{\theta}_0) = \exp\{(X_k\mu_k - \alpha_k(\mu_k))/\psi_k + b_k(X_k, \psi_k)\},$$

where $\boldsymbol{\theta}_k = (\theta_{k1}, \cdots, \theta_{kd})$ are the parameters specific to the treatment k, $\boldsymbol{\theta}_0 = (\theta_{01}, \cdots, \theta_{0d_0})$ are the common parameters for all groups. Assume the scale parameter ψ_k is fixed. Let $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i0}, \boldsymbol{Z}_{is})$. $\boldsymbol{Z}_{i0}$ are the covariates corresponding to the common parameters, and $\boldsymbol{Z}_{is}$ are the covariates corresponding to the group specific parameters. The link function is $\mu_k = h_k(\boldsymbol{Z}_{i0}\boldsymbol{\theta}_0' + \boldsymbol{Z}_{is}\boldsymbol{\theta}_k' I(T_{i,k} = 1))$. Simulation studies

for testing interaction effect in two groups were run to compare the performance of Zhang et al.'s (2007) CARA, complete randomization, and the proposed procedure. All covariates are assumed to have a Bernoulli distribution. For the two CARA procedures, 10 percent of the total sample size is assigned in the initial stage with equal allocation. The allocation rule for the proposed procedure is given by

$$\pi_1 = \Phi\left(\frac{\hat{\theta}_{11}(i) + \hat{\theta}_{12}(i)Z_{(i+1)2} - \hat{\theta}_{21}(i) - \hat{\theta}_{22}(i)Z_{(i+1)2}}{\gamma}\right).$$

The same allocation rule is used for Zhang et al.'s (2007) procedure:

$$\pi_1 = \Phi\left(\frac{\hat{\theta}^1_{01}(i)Z_{(i+1)1} + \hat{\theta}_{11}(i) + \hat{\theta}_{12}(i)Z_{(i+1)2} - \hat{\theta}^2_{01}(i)Z_{(i+1)1} - \hat{\theta}_{21}(i) - \hat{\theta}_{22}(i)Z_{(i+1)2}}{\gamma}\right), \tag{1.1}$$

where γ is the tuning parameter to set the desired skewing proportion. Compared with Zhang et al.'s (2007) procedure, the proposed procedure demonstrates higher power in linear and logistic regression.

1.2.3 Weighted Optimality Approach

Zhang and Hu (2009) proposed a covariate-adjusted doubly adaptive biased coin design. Write

$$\hat{\rho}_m = \frac{\sum_{i=1}^{m} \pi_1(\hat{\boldsymbol{\theta}}_m, \boldsymbol{Z}_i)}{m}$$

and

$$\hat{\pi}_m = \pi_1(\hat{\boldsymbol{\theta}}_m, \boldsymbol{Z}_{m+1}).$$

The randomization function to assign the $(m+1)$th subject to treatment 1 is given

by

$$\phi_{m+1,1} = \frac{\hat{\pi}_m \left(\frac{\hat{\rho}_m}{N_{m1}/m}\right)^{\gamma}}{\hat{\pi}_m \left(\frac{\hat{\rho}_m}{N_{m1}/m}\right)^{\gamma} + (1-\hat{\pi}_m)\left(\frac{1-\hat{\rho}_m}{1-N_{m1}/m}\right)^{\gamma}},$$

where $\gamma \geq 0$ is a constant controlling the degree of randomness. The asymptotic properties are derived. Zhang et al.'s (2007) design is a special case of Zhang and Hu's (2009) with $\gamma = 0$ which has the largest variability.

Biswas et al. (2012) developed a optimal CARA procedure using the log odds ratio for two-group longitudinal binary outcomes within Bayesian framework. Treatment-by-covariate interactions are not considered in this procedure. Different covariance correlation structures (constant correlation, zero correlation, AR(1), and AR(2) type structures) are considered in the simulation studies. The proposed procedures are compared to covariate incorporated longitudinal randomized play the winner design. The proposed procedures assign more patients to the better treatment group. The correlation between responses do not impact the allocation proportions, but affect the testing power. Misspecification of the correlation matrix significantly decrease the allocation proportions toward the better treatment.

Antognini and Zagoraiou (2012) described optimal designs for inference and ethics issues and proposed the reinforced doubly-adaptive biased coin design that included both continuous and discontinuous randomization functions. Simulation studies were run to compare the proposed continuous randomization with Zhang et al. (2007) CARA procedure and the discontinuous procedure to covariates-adjusted version of

Hu et al.'s (2009) efficient randomized adaptive design (ERADE), which is defined as

$$\phi_{ERADE,m+1} = \begin{cases} \alpha\hat{\pi}_{m,1}, & \text{if } N_{m,1}/m > \hat{\pi}_{m,1} \\ \hat{\pi}_{m,1}, & \text{if } N_{m,1}/m = \hat{\pi}_{m,1} \\ 1-\alpha(1-\hat{\pi}_{m,1}) & \text{if } N_{m,1}/m < \hat{\pi}_{m,1}, \end{cases}$$

where the constant $\alpha \in [0, 1)$ controls the degree of randomness. Compared with Zhang et al. (2007) CARA and extended ERADE procedures, both proposed procedures balance the variability of the allocation proportions better among different population strata. In general, discontinuous randomization functions perform better than continuous ones.

Sverdlov et al. (2013) proposed CARA randomization procedures for survival outcomes with two treatment groups when the outcome follows an exponential regression model. They used two approaches for a survival trial: CARA randomization procedures with a target and weighted optimality CARA randomization procedures. The simulation studies compare two balanced procedures, six CARA procedures, and two response-adaptive randomization procedures. They find that the proposed CARA procedures have similar power and type I error rates, fewer events compared with the balanced randomization procedures, and are robust to model misspecification. Delayed responses have a significant impact on convergence to the target allocations for all CARA and response-adaptive procedures. The ethical gains of CARA procedures with delayed responses are smaller than in the case of no delay.

Chang and Park (2013) used Bandyopadhyay et al.'s (2007) design and proposed a sequential estimation scheme. The proposed sequential estimation is based on a martingale estimating equation and the stopping rule depends on the observed Fisher

information. They show that, under the proposed stopping function, the asymptotic properties of the allocation function are the same as those in the non-sequential scenario. Simulation studies were conducted with binary responses, two treatment groups, and one continuous covariate. They find that the stopping time is very unstable when the initial sample size is too small. Compared to complete randomization with the same stopping rule, most of the CARA designs allocate more responsive patients to the better treatment group. Since the stopping rule is based on the estimate of the minimum eigenvalue, the required sample size increases as the number of covariates increasing. Highly correlated covariates also require a larger sample size.

Cheung et al. (2014) pointed out that Zhang et al.'s (2007) work does not consider the distribution theorems on the estimation of parameters in a reduced model. They developed the theorems needed and ran simulations under a logistic regression model. The treatment effect was tested by the likelihood ratio test. Three CARA procedures with different allocation rules and the complete randomization procedure were compared: CARA1 with allocation rule as $\pi_1 = (p_1/q_1)/(p_1/q_1 + p_2/q_2)$; CARA2 with allocation rule as $\pi_2 = \sqrt{p_1}/(\sqrt{p_1} + \sqrt{p_2})$; CARA3 with allocation rule as $\pi_3 = (q_2\sqrt{p_2})/(q_1\sqrt{p_1} + q_2\sqrt{p_2})$. All CARA procedures have lower failure rates and power comparable to the complete randomization procedure. The CARA2 procedure has the highest power and the lowest success rate among three CARA procedures. The CARA3 procedure has slightly higher power than CARA1 when there is no treatment-by-covariate interaction effect.

Hu et al. (2015) proposed CARA procedures based on efficiency and ethics for two treatment groups. They denoted the efficiency and ethics measurements of the two treatments as $d(Z, \theta) = (d_1(Z, \theta), d_2(Z, \theta))$ and $e(Z, \theta) = (e_1(Z, \theta), e_2(Z, \theta))$. The

randomization function to assign the $(m+1)$th subject to treatment 1 is given by

$$\pi_{m+1}(Z_{m+1}, \hat{\theta}(m)) = \frac{e_1(Z_{m+1}, \hat{\theta}(m)) d_1^{\gamma}(Z_{m+1}, \hat{\theta}(m))}{e_1(Z_{m+1}, \hat{\theta}(m)) d_1^{\gamma}(Z_{m+1}, \hat{\theta}(m)) + e_2(Z_{m+1}, \hat{\theta}(m)) d_2^{\gamma}(Z_{m+1}, \hat{\theta}(m))},$$

where $\gamma \geq 0$ is a tuning parameter that balanced the efficiency and ethics components. They used the $D-$optimality criteria as the efficiency measurement d_k and study various choices of e_k and γ. The simulation studies were performed for binary and normal covariates and outcomes. The design was compared with the complete randomization design and other three adaptive design. It performs better in case of type I error rate, power, and success rate under different choices of e_k and γ. Type I error rate and power are found based on the likelihood ratio test on the differential covariate effect (interaction effect).

Biswas and Bhattacharya (2016) considered the location and variability of the response distribution for CARA randomization. For a two groups clinical trial with continuous treatment outcomes, the proposed randomization function is given by

$$\pi_{\gamma}^{A}(\theta_A, \theta_B, z) = (\gamma_1 - \bar{\gamma}_2) P(Y_B - Y_A < \Delta(z)|z) + (\gamma_1 - \gamma_2) P(Y_A^2 < Y_B^2|z) + \bar{\gamma}_1,$$

where $\Delta(z) = \mu_A(z) - \mu_B(z)$, $\bar{\gamma}_1 = 1 - \gamma_1, \bar{\gamma}_2 = 1 - \gamma_2$. γ_1 and γ_2 are the weights to balance between ethics and variability. For the continuous outcome $X_A(X_B)$, $Y_A(Y_B)$ is defined as the difference between $X_A(X_B)$ and the mean $\mu_A(\mu_B)$. Simulation studies with different parameter values show that the proposed procedure not only assigns a higher proportion of subjects to the better treatment group, but also detects a small departure in treatment effectiveness with high probability, although a loss of power is also observed.

1.2.4 Bayesian Adaptive Randomization Methods

Bayesian adaptive randomization procedures are selection designs. Research has been done for binary, continuous, and survival outcomes. Biswas and Angers (2002) presented a Bayesian formulation of an adaptive design for clinical trial with continuous responses, two treatment groups, some prognostic covariate factors. Cheung et al. (2006) proposed exact and approximate Bayesian response-adaptive randomization procedures with and without covariate adjustment based on suvival outcomes. The simulations show that the approximate Bayes method with covariate adjustment seems to be robust to link misspecification. They assume that there is no treatment-by-covariate interaction effect. Yuan et al. (2011) proposed a Bayesian response-adaptive covariate-balanced randomization procedure for multi-arm clinical trials. The idea is to incorporate a covariate-adaptive randomization scheme into a Bayesian response-adaptive randomization. The updating of the posterior mean of the estimated parameters can be done continuously after each patient or after each group of patients. The simulation studies are run under the logistic regression model with two treatment groups and three covariates. The proposed procedure is compared with a procedure with equal allocation and three other response-adaptive or covariate-adaptive procedures. The proposed procedure successfully skews the allocation probablity to superior treatment group like response-adaptive procedures and has similar performance in balancing the covariates like covariate-adaptive procedures.

1.2.5 Non-parametric CARA Procedures

Bandyopadhyay and Bhattacharya (2012) developed an urn-based CARA procedure for binary responses trials with ordinal covariates. They compared the proposed

randomization with the stratified randomized play-the-winner rule in a hypothetical clinical trial for testing the treatment effect between groups with and without treatment-by-covariate interaction. Both the urn-based CARA and the stratified randomized play-the-winner procedures skew the randomization probabilities toward the desired direction. The urn-based CARA procedure has higher randomization probabilities with no more than 2 percent loss in power.

Aletti et al. (2018) proposed a class of CARA designs based on a new functinal urn model. This class of designs only requires independent but non-identically distributed covariates for all patients. The distribution of the responses conditioned on covariates is estimated nonparametrically. The urn is represented by a multivariate function of covariates and each patient is assigned by sampling from the urn given his or her own covariate profile. The entire functional urn composition will be updated after each allocation. In the context of precision medicine, this allow the investigation to choose optimal treatments based on the covariate model, even when there is insufficient information for a particular covariate profile

1.3 Properties of the GLM Approach to CARA

Zhang et al. (2007) lay out a framework for general CARA procedures for $K(\geq 2)$ treatment groups. The asymptotic properties are studied and apply to generalized linear models (GLM).

1.3.1 General CARA Procedure Framework and Asymptotic Properties

Based on the notation in Section 1.1, assume a patient with covariate vector $\boldsymbol{Z}$ is assigned to treatment $k, k = 1, \cdots, K$, the observed response is X_k, and the response

and covariate vector satisfy:

$$E(X_k|Z) = p_k(\theta_k, Z),\ \theta_k \in \Theta_k,\ k = 1, \cdots, K,$$

where $p_k(.,.),\ k = 1, \cdots, K$, are some known functions. $\theta_k,\ k = 1, \cdots, K$, are unknown parameters, and $\Theta_k \subset \mathbb{R}^d$ is the parameter space of θ_k. Denote $\theta = (\theta_1, \cdots, \theta_K)$ and $\Theta = \Theta_1 \times \cdots \times \Theta_K$. As in Section 1.1 , let $\boldsymbol{T}_m = (T_{m1}, \cdots, T_{mK})$ be the treatment assignment of the $m-$th patient. $\{X_{mk}, k = 1, \cdots, K,\ m = 1, 2, \cdots\}$ be the responses and $\{\boldsymbol{Z}_m, m = 1, 2, \cdots\}$ be the corresponding covariates. Assume that $\{X_{m1}, \cdots, X_{mK}, Z_m), m = 1, 2, \cdots\}$ is a sequence of i.i.d. random vectors, and the distribution s are the same as $(X_1, \cdots, x_K, \boldsymbol{Z})$.

The CARA procedure starts with assigning m_0 subjects to each treatment through restricted randomization. Assume that $m(m \geq Km_0)$ subjects have been assigned to treatments. The responses $\{\boldsymbol{X}_j, j = 1, \cdots, m\}$ and corresponding covariates $\{\boldsymbol{Z}_j, j = 1, \cdots, m\}$ are observed. Let $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_m)$ be the estimator of $\theta = (\theta_1, \cdots, \theta_m)$. For each $k = 1, \cdots, K,\ \hat{\theta}_{mk} = \hat{\theta}_{mk}(X_{jk}, Z_j : T_{jk} = 1, j = 1, \cdots, m)$ is the estimator of θ_k based on observed N_{mk} - sized sample (X_{jk}, Z_j). We then assign the $(m+1)$th subject with covariate Z_{m+1} to treatment k with a probability of :

$$\phi_{m+1,k} = E(T_{m+1,k}|\mathcal{F}_m, \mathcal{Z}_{m+1}) = E(T_{m+1,k}|\mathcal{X}_m, \mathcal{T}_m, \mathcal{Z}_{m+1}) = \pi_k(\hat{\theta}_m, Z_{m+1}), \quad (1.2)$$

$k = 1, \cdots, K$, where $\pi_k(\cdot,\cdot), k = 1, \cdots, K$ are some given functions. Given $\mathcal{F}_m$ and $\mathcal{Z}_{m+1}$, the response X_{m+1} of the $(m+1)$-th subject is assumed to be independent of its assignment T_{m+1} . Define $\pi(\cdot,\cdot) = \pi_1(\cdot,\cdot), \cdots, \pi_K(\cdot,\cdot))$ to be the randomization function that satisfies $\pi_1 + \cdots + \pi_K \equiv 1$. Let $g_k(\theta^*) = E[\pi_k(\theta^*, Z)]$, from (1.3.1), it

follows that

$$Pr(T_{m+1,k} = 1|\mathcal{X}_m, \mathcal{T}_m, \mathcal{Z}_m) = g_k(\hat{\theta}_m),\ k = 1, \cdots, K. \tag{1.3}$$

Different choices of $\pi(\cdot, \cdot)$ generate different class of randomization functions. We can take $\pi_k(\boldsymbol{\theta}, \boldsymbol{Z}) = R_k(\boldsymbol{\theta}_1\boldsymbol{Z}', \cdots, \boldsymbol{\theta}_K\boldsymbol{Z}'), k = 1, \cdots, K$. Here $0 < R_k(\boldsymbol{z}) < 1, k = 1, \cdots, K$ are real functions that are defined on $\mathbb{R}^K$ with

$$\sum_{k=1}^{K} R_j(\boldsymbol{z}) = 1, \text{and } R_i(\boldsymbol{z}) = R_j(\boldsymbol{z}) \text{ whenever} z_i = z_j. \tag{1.4}$$

Assume that $\boldsymbol{Z}$ and $\boldsymbol{\theta}_k, k = 1, \cdots, K$ have the same dimensions. In practice, the functions R_k can be defined as

$$R_k(\boldsymbol{z}) = \frac{G(z_k)}{G(z_1) + \cdots + G(z_K)}, \quad k = 1, \cdots, K,$$

where G is a smooth positive real function that is defined in $\mathbb{R}$.

Define $g(\theta^*) = (g_1(\theta^*), \cdots, g_K(\theta^*))$ and let $v_k = g_k(\boldsymbol{\theta}) = E[\pi_k(\boldsymbol{\theta}, \boldsymbol{Z})], k = 1, \cdots, K$ and $v = (v_1, \cdots, v_K)$. Assume $0 < v_k < 1, k = 1, \cdots, K$. For the randomization function $\pi(\boldsymbol{\theta}^*, \boldsymbol{z})$, we assume the following conditions:

Condition A. Assume that parameter space Θ_k is a bounded domain in $\mathbb{R}^d$, and the true value $\boldsymbol{\theta}_k$ is an interior point of $\Theta_k, k = 1, \cdots, K$. For each fixed $\boldsymbol{z}, \pi_k(\theta^*, \boldsymbol{z}) > 0$ is a continuous function of $\theta^*, k = 1, \cdots, K$; for each $k = 1, \cdots, K, \pi_k(\theta^*, \boldsymbol{Z})$ is differentiable with respect to θ^* under the expectation, and there exists a $\delta > 0$ such that

$$g_k(\boldsymbol{\theta}^*) = g_k(\boldsymbol{\theta}) + (\boldsymbol{\theta}^* - \boldsymbol{\theta})(\frac{\partial g_k}{\partial \boldsymbol{\theta}^*}\ |_\theta)^T + o(||\boldsymbol{\theta}^* - \boldsymbol{\theta}||^{1+\delta}),$$

where $\partial g_k/\partial \boldsymbol{\theta}^* = (\partial g_k/\partial \theta^*_{11}, \cdots, \partial g_k/\partial \theta^*_{Kd})$.

Condition B. Suppose that for $k = 1, \cdots, K$,

$$\hat{\boldsymbol{\theta}}_{nk} - \boldsymbol{\theta}_k = \frac{1}{n}\sum_{m=1}^{n} T_{mk} h_k(X_{mk}, \boldsymbol{Z}_m)(1 + o(1)) + o(n^{-1/2}) \quad a.s. \tag{1.5}$$

where h_k are K functions with $E[h_k(X_k, \boldsymbol{Z})|\boldsymbol{Z}] = 0, k = 1, \cdots, K$.

THEOREM 1. If $E||h_k(X_k, \boldsymbol{Z})||^{2+\epsilon} < \infty$ for some $\epsilon > 0,\ k = 1, \cdots, K$, then under condition A and B, we have for $k = 1, \cdots, K$,

$$Pr(T_{nk} = 1) \to v_k;\ Pr(T_{nk} = 1|\mathcal{F}_{n-1}, \boldsymbol{Z}_n = \boldsymbol{z}) \to \pi_k(\boldsymbol{\theta}, \boldsymbol{z})\ a.s. \tag{1.6}$$

and

$$\frac{N_n}{n} - v = O\left(\sqrt{\frac{\log\log n}{n}}\right)\ a.s.; \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} = O\left(\sqrt{\frac{\log\log n}{n}}\right). \tag{1.7}$$

Further, let $\boldsymbol{V} = diag(\boldsymbol{V}_1, \cdots, \boldsymbol{V}_K)$, where

$$\boldsymbol{V}_k = E\{\pi_k(\boldsymbol{\theta}, \boldsymbol{Z})(h_k(X_k, \boldsymbol{Z}))' h_k(X_k, \boldsymbol{Z})\},\ k = 1, \cdots, K,$$

$$\boldsymbol{\Sigma}_1 = diag(\boldsymbol{v}) - \boldsymbol{v}'\boldsymbol{v}, \quad \boldsymbol{\Sigma}_2 = \sum_{k=1}^{K} \frac{\partial g}{\partial \theta_k} \boldsymbol{V}_k \left(\frac{\partial g}{\partial \theta_k}\right)', \text{and } \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + 2\boldsymbol{\Sigma}_2.$$

Then,

$$\sqrt{n}(N_n/n - v) \to N(\boldsymbol{0}, \boldsymbol{\Sigma}) \text{ and } \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \to N(\boldsymbol{0}, \boldsymbol{V}) \tag{1.8}$$

in distribution.

Theorem 1 gives the asymptotic properties of the overall allocation proportions N_n/n. Sometimes, we may want to know the allocation proportions for a given set

of covariates. For a given set of covariate vector $\boldsymbol{z}$, the allocation proportion to treatment k is :

$$\frac{\sum_{m=1}^{n} T_{mk} I\{\boldsymbol{Z}_m = \boldsymbol{z}\}}{\sum_{m=1}^{n} I\{\boldsymbol{Z}_m = \boldsymbol{z}\}} := \frac{N_{n,k|\boldsymbol{z}}}{N_n(\boldsymbol{z})},$$

where $N_{n,k|\boldsymbol{z}}$ is the number of subjects with covariate $\boldsymbol{z}$ that is randomized to treatment k, $k = 1, \cdots, K$ in n trials. Let $N_n(\boldsymbol{z})$ be the total number of subjects with covariate $\boldsymbol{z}$ and let $N_{n|\boldsymbol{z}} = (N_{n,1|\boldsymbol{z}}, \cdots, N_{n,K|\boldsymbol{z}})$. Theorem 2 below gives the asymptotic properties of the conditional proportions.

THEOREM 2. Given a set of covariates $\boldsymbol{z}$, assume that $Pr(\boldsymbol{Z} = \boldsymbol{z}) > 0$. Under conditions A and B, we have

$$N_{n,k|\boldsymbol{z}}/N_n(\boldsymbol{z}) \to \pi_k(\boldsymbol{\theta}, \boldsymbol{z}) \quad a.s. \quad k = 1, \cdots, K, \tag{1.9}$$

and

$$\sqrt{N_n(\boldsymbol{z})}(N_{n|\boldsymbol{z}}/N_n(\boldsymbol{z}) - \boldsymbol{\pi}(\boldsymbol{\theta}, \boldsymbol{z})) \to N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{z}}), \tag{1.10}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{z}} = diag(\boldsymbol{\pi}(\boldsymbol{\theta}, \boldsymbol{z})) - \boldsymbol{\pi}(\boldsymbol{\theta}, \boldsymbol{z})'\boldsymbol{\pi}(\boldsymbol{\theta}, \boldsymbol{z}) + 2\sum_{k=1}^{K} \frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta}, \boldsymbol{z})}{\partial \boldsymbol{\theta}_k} \boldsymbol{V}_k \left(\frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta}, \boldsymbol{z})}{\partial \boldsymbol{\theta}_k} \right)' Pr(\boldsymbol{Z} = \boldsymbol{z}).$$

Detailed proofs of Theorem 1 and Theorem 2 can be found in the Appendix of Zhang et al. (2007).

1.3.2 Generalized Linear Model

The general results in Section 1.3.1 can be applied to the generalized linear model. Suppose for a given covariate $\boldsymbol{Z}$, the response X_k of the treatment $k = 1, \cdots, K$ has

a distribution in the exponential family under the GLM:

$$f_k(X_k|\boldsymbol{Z},\boldsymbol{\theta}_k) = \exp\{(X_k\mu_k - \alpha_k(\mu_k))/\psi_k + b_k(X_k,\psi_k)\}, \tag{1.11}$$

with the link function $\mu_k = h_k(\boldsymbol{Z}\boldsymbol{\theta}_k')$, where $\theta_k = (\theta_{k1},\cdots,\theta_{kd}),\ k = 1,\cdots,K$. Assume the scale parameter ψ_k is fixed. Under this model, $E(X_k|Z) = \alpha'(\mu_k),\ \mathrm{Var}(X_k|Z) = \alpha''(\mu_k)\psi_k$. The first and second derivatives for log likelihood function are:

$$\frac{\partial log f_k(X_k|\boldsymbol{Z},\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k} = \frac{1}{\psi_k}(X_k - a_k'(\mu_k))h_k'(\boldsymbol{Z}\boldsymbol{\theta}_k')\boldsymbol{Z}$$

and

$$\frac{\partial^2 log f_k(X_k|\boldsymbol{Z},\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k^2} = \frac{1}{\psi_k}\{-a_k''(\mu_k))[h_k'(\boldsymbol{Z}\boldsymbol{\theta}_k')]^2 + [X_k - a_k'(\mu_k)]h_k''(\boldsymbol{Z}\boldsymbol{\theta}_k')\}\boldsymbol{Z}'\boldsymbol{Z}.$$

For a given covariate $\boldsymbol{Z}$, the conditional Fisher's information matrix is given by

$$\boldsymbol{I}_k(\boldsymbol{\theta}_k|\boldsymbol{Z}) = -E\left[\frac{\partial^2 log f_k(X_k|\boldsymbol{Z},\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k^2}\middle|\boldsymbol{Z}\right] = \frac{1}{\psi_k}a_k''(\mu_k)[h_k'(\boldsymbol{Z}\boldsymbol{\theta}_k')]^2\boldsymbol{Z}'\boldsymbol{Z}.$$

For the observations up tp m, the likelihood function is given by

$$\boldsymbol{L}(\boldsymbol{\theta}) = \prod_{j=1}^{m}\prod_{k=1}^{K}[f_k(X_{jk}|\boldsymbol{Z}_j,\boldsymbol{\theta}_k)]^{T_{jk}} = \prod_{k=1}^{K}\prod_{j=1}^{m}[f_k(X_{jk}|\boldsymbol{Z}_j,\boldsymbol{\theta}_k)]^{T_{jk}} := \prod_{k=1}^{K}\boldsymbol{L}_k(\boldsymbol{\theta}_k),$$

where $\log \boldsymbol{L}_k(\boldsymbol{\theta}_k) \propto \sum_{j=1}^{m} T_{jk}(X_jk - a_k(\mu_{jk})),\ \mu_{jk} = h_k(\boldsymbol{\theta}_k'\boldsymbol{Z}_j), k = 1,\cdots,K$. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_m = (\hat{\boldsymbol{\theta}}_{m,1},\cdots,\hat{\boldsymbol{\theta}}_{m,K})$ maximizes $\boldsymbol{L}(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1 \times$

$\cdots \times \boldsymbol{\Theta}_K$. And $\hat{\boldsymbol{\theta}}_{mk}$ maximizes L_k over $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k,\ k = 1, \cdots, K$.

Corollary 1. Let $v_k = E[\pi_k(\boldsymbol{\theta}, \boldsymbol{Z})], \boldsymbol{I}_k = \boldsymbol{I}_k(\boldsymbol{\theta}) = E[\pi_k(\boldsymbol{\theta}, \boldsymbol{Z})\boldsymbol{I}_k(\boldsymbol{\theta}_k|\boldsymbol{Z})],\ k = 1, \cdots, K$. Suppose that a_k'', h_k'' are continuous, $\boldsymbol{Z}$ is bounded, matrices $\boldsymbol{I}_k,\ k = 1, \cdots, K$, are nonsingular, and the MLE $\hat{\boldsymbol{\theta}}_m$ is unique, then under Condition A, we have (1.3.5), (1.3.6), and (1.3.7) with $\boldsymbol{V}_k = \boldsymbol{I}_k^{-1},\ k = 1, \cdots, K$; and if $Pr(\boldsymbol{Z} = \boldsymbol{z}) > 0$ for a given covariate $\boldsymbol{z}$, then (1.3.8) and (1.3.9) also hold. For logistic regression model, we have $\boldsymbol{I}_k = E[\pi_k(\boldsymbol{\theta}, \boldsymbol{Z})p_k q_k \boldsymbol{Z}'\boldsymbol{Z}],\ k = 1, \cdots, K$. For normal linear regression model, we have $\boldsymbol{I}_k = E[\pi_k(\boldsymbol{\theta}, \boldsymbol{Z})\boldsymbol{Z}'\boldsymbol{Z}]/\sigma_k^2,\ k = 1, \cdots, K$

When the distribution of $\boldsymbol{Z}$ and the true value of $\boldsymbol{\theta}$ are unknown, we obtain the estimates as follows:

(a) Estimate $\boldsymbol{I}_k$ by $\hat{\boldsymbol{I}}_{nk} = \left(\sum_{m=1}^{n} T_{nk}\right)^{-1} \sum_{m=1}^{n} T_{nk}\boldsymbol{I}_k(\hat{\boldsymbol{\theta}}_{nk}|\boldsymbol{Z}_m),\ k = 1, \cdots, K$, and $\boldsymbol{V}$ by $\hat{\boldsymbol{V}}_n = diag(\hat{\boldsymbol{I}}_{n1}^{-1}, \cdots, \hat{\boldsymbol{I}}_{nK}^{-1})$.

(b) Estimate $\boldsymbol{\Sigma}_1$ by $\hat{\boldsymbol{\Sigma}_1} = diag\left(\frac{N_n}{n}\right) - \left(\frac{N_n}{n}\right)'\frac{N_n}{n}$, and $\frac{\partial g}{\partial \boldsymbol{\theta}_k}$ by $\frac{\hat{\partial g}}{\partial \boldsymbol{\theta}_k} = \frac{1}{n}\sum_{m=1}^{n} \frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta}^*, \boldsymbol{Z}_m)}{\partial \boldsymbol{\theta}_k^8}|_{\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}_n}$

(c) Estimate $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}_1} + 2\sum_{k=1}^{K} \frac{\hat{\partial g}}{\partial \boldsymbol{\theta}_k}\hat{\boldsymbol{V}}_k\left(\frac{\hat{\partial g}}{\partial \boldsymbol{\theta}_k}\right)'$

(d) For a given covariate $\boldsymbol{z}$, $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ is estimated by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = diag(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{z})) - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{z})'\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{z}) + 2\sum_{k=1}^{K}\left(\frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta}^*, \boldsymbol{z})}{\partial \boldsymbol{\theta}_k^*}|_{\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}_n}\right)\hat{\boldsymbol{V}}_k\left(\frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta}^*, \boldsymbol{z})}{\partial \boldsymbol{\theta}_k^*}|_{\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}_n}\right)' \frac{\#\{m \le n : \boldsymbol{Z}_m = \boldsymbol{z}\}}{n}$.

To test the homogeneity among the treatments, that is,

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_K \text{ versus } H_1 : \text{ not all } \boldsymbol{\theta}_k \text{ are equal.}$$

We define

$$\boldsymbol{\theta}^c = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_K, \cdots, \boldsymbol{\theta}_{K-1} - \boldsymbol{\theta}_K),\ \hat{\boldsymbol{\theta}}^c = (\hat{\boldsymbol{\theta}}_{n,1} - \hat{\boldsymbol{\theta}}_{n,K}, \cdots, \hat{\boldsymbol{\theta}}_{n,K-1} - \hat{\boldsymbol{\theta}}_{n,K})$$

and

$$\boldsymbol{V}^c = diag(\boldsymbol{I}_1^{-1}, \cdots, \boldsymbol{I}_{K-1}^{-1}) + 1'1 \otimes \boldsymbol{I}_K^{-1},\ \hat{\boldsymbol{V}}^c = diag(\hat{\boldsymbol{I}}_1^{-1}, \cdots, \hat{\boldsymbol{I}}_{K-1}^{-1}) + 1'1 \otimes \hat{\boldsymbol{I}}_K^{-1}.$$

By (1.3.7), we have $\sqrt{n}(\hat{\boldsymbol{\theta}}^c - \boldsymbol{\theta}^c) \to N(\mathbf{0}, \boldsymbol{V}^c)$ in distribution. Therefore, for testing the homogeneity of the treatments, a natural test statistic is $n\hat{\boldsymbol{\theta}}^c(\hat{\boldsymbol{V}}^c)^{-1}(\hat{\boldsymbol{\theta}}^c)'$. The asymptotic distribution is $\mathcal{X}^2_{(K-1)d}$ under H_0 and $\mathcal{X}^2_{(K-1)d}(\varphi)$ under H_1 with the non-centrality parameter $\varphi = n\boldsymbol{\theta}^c(\boldsymbol{V}^c)^{-1}(\boldsymbol{\theta}^c)'$. This test is asymptotically equivalent to the likelihood ratio test.

1.4 General Definition of Adaptive Enrichment Design

With the rapid development in genomic and genetic research, precision medicine has gained more attention in modern clinical trials. Molecularly targeted therapies are likely to only work with a subgroup of biomarker-positive patients. Many clinical trial procedures have been developed to incorporate the biomarkers. An enrichment design, also called a targeted design, was first studied by Simon and Maitournam (2004), Maitournam and Simon (2005). In this single stage design, patients are screened and selected by their biomarker status such that only biomarker-positive patients are enrolled and randomized to treatment groups. Only one null hypothesis of no treatment effect in the biomarker-positive subgroup can be tested.

An adaptive design is defined as a multistage study design that uses accumulating data to decide how to modify aspects of the study without undermining the validity and integrity of the trial (Dragalin, 2006). In order to expand the testable hypotheses and/or deal with no clearly defined subgroup at the beginning of the trial, several adaptive enrichment designs have been proposed and studied. One of the first biomarker-based, adaptive enrichment designs was introduced by Wang et al. (2007). The proposed two-stage adaptive enrichment design randomizes all subjects to treatment or control groups in stage I. If the treatment effect reaches a futility boundary in the biomarker-negative group at the interim analysis, the recruitment of the biomarker-negative subjects is terminated at the second stage, and the remaining sample size is re-allocated to biomarker-positive patients. In this case, the primary hypothesis is to test the treatment effect in the biomarker-positive subgroup. Otherwise, if the futility threshold is not reached in the biomarker-negative group at the interim analysis, the trial continue for all patients, and both overall and subgroup-specific tests are performed. The sample size is calculated based on the non-adaptive approach and kept unchanged at the interim analysis.

However, in practice, there may not be a clearly defined subgroup at the beginning of the phase III trial. The biomarker may be continuous with no known cutpoint or no clear single biomarker available to define a subgroup. Simon and Simon (2013) proposed a phase III adaptive enrichment design which begins with all patients in the trial, and sequentially restricts entry in an adaptive manner. This enrichment approach does not require a predefined subgroup. The primary null hypothesis is that no subgroup benefits more from treatment over control. Adaptive enrichment designs may largely increase test power, especially when only a small subset of patients benefit from the treatment.

1.5 Adaptive Enrichment Designs in the Literature

Simon (2015) discusses the application of and challenges for adaptive enrichment designs under three common scenarios: a single categorical biomarker, a single continuous biomarker with unknown cut point, and multidimensional biomarkers/combining multiple candidate biomarkers. Strata-based designs are effective for simple categorical biomarkers, but they often cannot control type I error rate well. For univariate continuous biomarkers, strata-based designs need a pre-defined strata and not leverage the ordering of the categories; model-based designs do not require pre-defined strata, but they can only test a single null hypothesis of no subgroup benefits more from treatment over control. Model-based designs are more effective for multivariate biomarkers.

1.5.1 Strata-Based Adaptive Enrichment Design

Strata-based adaptive enrichment designs have clearly defined subgroup characteristics at the design stage. Russek-Cohen and Simon (1997) proposed a two-stage procedure to investigate whether males and females respond differently to treatments. The proposed procedure tests for a gender by treatment interaction at the first stage. If no significant interaction is found, the study will be terminated and an overall treatment effect will be computed. If a significant interaction is found, the probability of going to the second stage for one or both genders will be calculated. When a strong gender-by-treatment interaction exists, it is more likely that the gender which is not benefitting from the treatment will go to the second stage. The two-stage adaptive enrichment design proposed by Wang et al. (2007) randomizes all subjects to treatment or control groups in stage I. If the treatment effect reaches the futility boundary in the known

biomarker-negative group at the interim analysis, the recruitment of the biomarker-negative subjects is terminated at the second stage and the remaining sample size is re-allocated to biomarker-positive patients. In this case, the primary hypothesis is to test the treatment effect in the biomarker-positive subgroup. Otherwise, if the futility threshold is not reached in the biomarker-negative group at the interim analysis, the trial continues for all patients and both overall and subgroup-specific tests are performed. The sample size is calculated based on the non-adaptive approach and kept unchanged at the interim analysis. The performances of the proposed design with different α allocation rules are compared with the fixed design and Freidlin and Simon's (2005) adaptive signature design. The findings show that the proposed adaptive design outperforms the fixed design when the biomarker can predict which patient subgroup benefits from the treatment. In the latter work, Wang et al. (2009) expand the framework to nested patient subsets when multiple predefined categorical markers are presented.

Mehta and Gao (2011) proposed an adaptive enrichment design, where adaptive modification is made of an ongoing group sequential trial based on an interim analysis. The modifications include adaptations in the number, spacing, and information times of subsequent interim analyses, as well as population enrichment.

In order to evaluate the overall treatment effect in both biomarker-positive and biomarker-negative patients in enrichment designs, Yang et al. (2015) proposed an enrichment design with patient population augmentation. Specifically, after sufficiently powering the biomarker-positive subgroup, biomarker-negative patients are enrolled to assess the overall treatment benefit. A weighted statistic is used to correct for the disproportionality of biomarker-positive and biomarker-negative groups under the enriched trial setting. Screening is needed to obtain the information on the weight determination. Simulation results showed that the proposed design can

safeguard the power for biomarker-positive subgroup with a slightly larger sample size when there is misspecification on the treatment effect in the biomarker-negative subgroup at the design stage.

1.5.2 Model-Based Adaptive Enrichment Design

Model-based adaptive enrichment designs are two-stage designs where the complete specification of subgroup characteristics are only available at the end of the first stage, based on the interim analysis. Simon and Simon (2013) developed a very general model and statistical significance tests for eligibility modification. They illustrate the framework in the setting of adaptive threshold enrichment of a single continuous biomarker with no known cut-off at the beginning of the trial. Renfro et al. (2014) proposed a two-stage adaptive enrichment phase II design incorporating prospective continuous marker threshold selection, possible early futility stopping, possible mid-trial accrual restriction, and final marker and treatment evaluation in marker-positive patients. The proposed design assumes time-to-event endpoints, unequal allocation in the first stage, possible cutoff point within the range of 25% to 75%, and treatment effect in monotone non-decreasing function on the continuous biomarker variable. The cutoff is selected by minimizing the p-value of the interaction between the treatment effect and the dichotomized biomarker. Simulation studies demonstrated that type I error rates are in the acceptable range. The power is highly depend on the successful classification of the true predictive biomarker at the interim analysis. This critical classification of the biomarker depends not only on the biomarker prevalence and effect size, but also on the timing of the interim analysis.

Spencer et al. (2016) proposed a continuous biomarker-adaptive threshold trial design, which both selectively recruits from the start of the trial and also modifies the eligibility criteria to a targeted subgroup that will have a statistically significant

response rate. All subjects are used in the final test for efficacy, even if the eligibility criteria is changed at the interim. This design aims to demonstrate that there is a subgroup in which the treatment is effective and estimate the most appropriate value of the biomarker threshold to define the boundary of this subgroup. The basic study design contains a single arm, which is typical for early stage phase II oncology trials to test efficacy, but it can be modified to include a control arm. They describe a single-arm trial design for a treatment with a binary outcome and assume that the treatment effect is a monotone non-decreasing function on the continuous biomarker variable. The total and stage specific sample sizes are fixed before the study begins. A preliminary threshold is chosen based on prior knowledge at stage I and is updated at stage II based on the results from the first stage. The single continuous biomarker is converted to its estimated quantiles and assumed to follow uniform $(0, 1)$ distribution. A binomial exact test using all subjects recruited to the trial to test whether the response rate exceeds a pre-defined reference rate. Simulation studies comparing the adaptive design and the fixed design showed that both methods have conservative type I error rates in the overall simulated studies. The adaptive design has higher power in both overall and completed studies when the true threshold value is above the 0.3 quantile. The estimated bias of the threshold value can be obtained through the simulation studies even if the hypothesis test is non-significant.

Ohwada and Morita (2016) proposed a Bayesian adaptive design using a four-parameter change-point model to stop enrollment of insensitive patients at the interim analysis, in a setting of a phase II two-group randomized clinical trial with a time-to-event outcome and a single continuous biomarker. They also assume that the treatment effect is a monotone non-decreasing function on the continuous biomarker variable. Two or three interim analyses are planned. For the jth patient with biomarker

level of z_j, the proportional hazard model for hazard at time t is assumed:

$$ln(h_j(t|T_j, z_j)) = ln(H_0(t)) + T_j f(z_j),$$

where $h_0(t)$ is the baseline hazard function and $f(z_j)$ is a function representing the relationship between the biomarker level and treatment effect. A four-parameter change-point model is proposed as:

$$f(x) = \beta_1 I(z < \xi_1) + \left(\frac{\beta_2 - \beta_1}{\xi_2 - \xi_1}(z - \xi_1) + \beta_1 \right) I(\xi_1 \leq z \leq \xi_2) + \beta_2 I(z \geq \xi_2),$$

where β_1, β_2, ξ_1 and ξ_2 are parameters with constraints $\beta_1 > \beta_2$ and $\xi_1 > \xi_2$. Within a Bayesian framework, the model is updated using all accumulated data at the interim analysis and the final analysis. Markov chain Monte Carlo is used for posterior computation with the following assumed non-informative priors:

$$\beta_1 \sim N(0, 1000),$$

$$\delta = \beta_1 - \beta_2 \sim \Gamma(0.001, 0.001),$$

and ξ_1, ξ_2 follow a uniform distribution with a probability density function of $(\xi_U^* - \xi_L^*)^2/2$ when $\xi_L^* < \xi_1 < \xi_2 < \xi_U^*$ and 0 otherwise, where ξ_L^* and ξ_U^* are predefined lower and upper limits of the biomarker, respectively. Simulation studies demonstrated that, compared with the standard no restriction approach, the proposed approach reduces the number of enrolled patients from the insensitive subgroup, with mild reduction in the probability of reaching a correct decision and identifying the sensitive subgroup. Additionally, the non-enrichment four-parameter change-point model perform better over a wide range of simulation scenarios than a commonly

used dichotomization non-enrichment approach.

Diao et al. (2018) introduced a biomarker threshold adaptive design with survival outcomes. In the first stage, based on historical or pilot studies, some subgroups are identified such that patients in these subgroups benefit the most from the new treatment. In the second stage, only patients from the subgroups determined in the first stage are recruited and randomly allocated to the treatment or control group.

1.6 Simon's Adaptive Enrichment Design

Simon and Simon (2013) introduce a phase III adaptive enrichment design which begins without restricting entry and sequentially restricts entry based on candidate biomarkers.

1.6.1 Adaptive Enrichment Design for Two Group Binary Outcome

Based on the notation in Section 1.1, assume that we want to compare a new treatment with control and start with equiprobable allocation. Let $T_i = 1$ for the new treatment and $T_i = 0$ for control. Let $X_i = 1$ for response and $X_i = 0$ for non-response.

After the first m patients, the recruitment is restricted to those patients who will benefit from the treatment. Let $f(Z)$ be the map from the covariate space to $\{0, 1\}$:

$$f(Z) = I\{p_T(\boldsymbol{Z}) > p_C(\boldsymbol{Z})\}, \tag{1.12}$$

where $p_T(\boldsymbol{Z})$ and $p_C(\boldsymbol{Z})$ are the probabilities of response for a patient with covariate vector $\boldsymbol{Z}$ under treatment and control. For each patient i, let $\hat{f}_i(\boldsymbol{Z})$ be the estimate of

$f(\boldsymbol{Z})$. $\hat{f}_i(\boldsymbol{Z})$ is computed based on all previous $i-1$ patients' responses, assignments, and covariates information. After the first m patients, for each $i > m$, we find $\hat{f}_i(\boldsymbol{Z})$ from previous $i-1$ patients, the entry into the clinical trial will be restricted to patients with $\hat{f}_i(\boldsymbol{Z}) = 1$ only. This process is repeated until a total of n patients have been enrolled. The null hypothesis is that no subgroup benefits more from treatment than control:

$$p_T(\boldsymbol{Z}) = p_C(\boldsymbol{Z}), \text{ for all } \boldsymbol{Z}.$$

In order to preserve the type I error rate, the number of successes on the treatment plus the number of failures on the control is used as the test statistic. Under the null hypothesis, $S \sim \text{binomial}(n, 0.5)$. Regardless of the classification methods used for modifying enrollment criteria, comparing S with the tails of this binomial is a valid test that preserves the type I error rate.

Assume patients are accepted and randomized in pairs, one to each treatment arm, and the enrollment criteria $\hat{f}$ is updated no more frequently than after each pair, then the test statistic proposed above is equivalent to:

$$\tilde{S} = \sum_{i=1}^{n} \left(I\{X_{i,T} > X_{i,C}\} - I\{X_{i,T} < X_{i,C}\} \right), \tag{1.13}$$

where $X_{i,C}$ and $X_{i,T}$ are the outcomes for the control subject and treatment subject in the ith pair. For a pre-specified number u of untied pairs, under the null hypothesis, McNemars test is used to control the type I error rate.

1.6.2 Adaptive Threshold Enrichment Design

When a single candidate predictive biomarker is available but no cutpoint has been determined at the beginning of phase III trials, the method described above can be applied. Assume that the treatment effect $p_T(z) - p_C(z)$ for a patient with biomarker value z is either 0 or δ and that the treatment effect is monotone non-decreasing in z with a jump only at one of the candidate cutpoints. Let $\xi_1, \cdots, \xi_K$ be a discrete set of candidate cutpoints, $p_C(z) = p_0$ for all z, $p_T(z) = p_0$ for $z \leq \xi_k$, and $p_T(z) = p_1$ for $z > \xi_k$, where $p_0 \leq p_1$. At the interim analysis, the candidate cutpoint ξ_k at which the log-likelihood is maximized is taken as an estimate of the true cutpoint, z^* and subsequent accrual is restricted to patients with biomarker values greater than that ξ_k. Simulation studies with different choices of $p_0, p_1, K,$ and z^* show that the adaptive design using statistic (1.6.1) has much higher power than equiprobable procedure for most conditions. Type I error rates are preserved even when the response rates change from pre- to post-interim analysis. Compared with equiprobable approach, the adaptive enrichment design is most powerful when only a small subset of patients benefit.

1.6.3 Group Sequential Analysis

When a group sequential analysis is needed, there are other strategies that preserve the type I error rate. For each block k, let s_k be some statistic based on the data in that block and n_k be the sample size in the kth block. When the distribution of each s_k is known and independent of $\mathcal{F}_{k-1}$ under the null, we may choose any function G and construct a valid test that preserves the type I error rate.

For continuous data, the proposed adaptive t-test statistic is given by

$$\frac{1}{\sqrt{n}} \sum_{k \leq K} \sqrt{n_k} \left(\frac{\bar{x}_{(T,k)} - \bar{x}_{(C,k)}}{\sqrt{\hat{\sigma}^2_{(T,k)}/(n_{T,k}-1) + \hat{\sigma}^2_{(C,k)}/(n_{C,k}-1)}} \right), \tag{1.14}$$

where $\bar{x}_{(T,k)}, \bar{x}_{(C,k)}, \hat{\sigma}^2_{(T,k)}, \hat{\sigma}^2_{(C,k)}, n_{T,k}$, and $n_{C,k}$ denote the treatment and control sample means, variances, and sample sizes in the kth block, respectively. Under the null hypothesis, for each sufficiently large n_k, the test statistic is asymptotically standard normal distributed. Simon and Simon (2013) show that this adaptive t-test statistic has the same limiting distribution as regular t statistic for full-population alternatives (i.e. all subgroups have the same distribution under the null, and identical change under the alternative). A block MannWhitneyWilcoxon test could be used if asymptotic normality is not assumed. The test statistic is given by

$$u = \sum_{k \leq K} w_k u_k,$$

where u_k is the MannWhitney statistic for kth block and w_k is a predefined weight. Under the null, the ranking of variables within any block is equally probable.

For binary data, the proposed adaptive design test statistic is given by

$$z = \frac{1}{\sqrt{n/2}} \sum_{k \leq K} \sqrt{n_k/2} \left(\frac{\hat{p}_{(T,k)} - \hat{p}_{(C,k)}}{2\sqrt{\hat{p}_{(\text{pool},k)}(1 - \hat{p}_{(\text{pool},k)})/n_k}} \right), \tag{1.15}$$

where $\hat{p}_{(T,k)}$ and $\hat{p}_{(C,k)}$ are treatment and control sample success proportions in kth block, respectively, and $\hat{p}_{(\text{pool},k)} = (\hat{p}_{(T,k)} + \hat{p}_{(C,k)})/2$.

For survival data, let $\ell_k(\beta)$ be the log-likelihood of the Cox model for the kth

block where β is the coefficient for the treatment indicator, with the first and second derivatives ℓ'_k and ℓ''_k respectively. The proposed adaptive design test statistic is given by

$$S = \sum_k w_k \frac{\ell'_k(0)}{\sqrt{-\ell''_k(0)}}, \tag{1.16}$$

where w_k are pre-specified non-negative weights. This test statistic is asymptotically $N(0, W)$, where W is the sum of the squares of the weights.

1.7 Randomization Tests

Randomization not only promotes comparability among the study groups, but also provides a distributional assumption-free basis for statistical inference. Rosenberger and Lachin (2015) thoroughly discuss the concept and method of performing randomization tests for assessing whether a treatment has any effect on the responses of the n patients randomized in the study. The null hypothesis of a randomization test is that the treatment assignments are unrelated to the responses of the n patients randomized in the study. An appropriate measure of the treatment effect is then used as the test statistic, which could be a difference of means or proportions, a nonparametric rank test, or a covariate-adjusted treatment effect measure (Parhat et al., 2014). One calculates the two-sided p-value as the proportion of the more extreme test statistics from the reference set than the observed test statistic. Let Ω be the reference set, which is the set of all possible permutations of randomization sequences and the associated probabilities. Define S_l to be the test statistic for a sequence l, $l = 1, \cdots, \Omega$ and S_{obs} to be the observed test statistic. Let L record realizations of

specific randomization sequences. The p-value is given by

$$p = \sum_{l=1}^{\Omega} I(|S_l| \geq |S_{obs}|) Pr(L = l).$$

We use Monte Carlo simulation to calculate the two-sided p value estimator as

$$\hat{p} = \frac{\sum_{l=1}^{L_s} I(|S_l| \geq |S_{obs}|)}{L_s},$$

where L_s is the total number of realizations. This technique cannot be used to test an interaction directly since under the null hypothesis of no interaction effect, there still might be overall treatment effect.

Randomization tests have been studied as assumption-free alternatives or complements to the traditional population model-based analyses. Rosenberger et al. (2019) discussed the advantages of randomization tests and the application of the randomization tests in testing the primary outcome and sequential monitoring. Other studies have tested the treatment effect among groups using randomization tests (Galbete and Rosenberger, 2016, Parhat et al., 2014, Plamadeala et al., 2012). Parhat et al. (2014) show that, under model misspecification, randomization tests preserve the size and power well for generalized linear regression, survival, and longitudinal models, while population model-based tests have inflated type I error rates and reduced power. Still and White (1981) showed the application of the randomization test in assessing the interaction effect in the analysis of variance settings. However, the effect and application of using randomization-based tests to examine the interaction effect in the generalized linear model setting when we have multiple covariates strata or categorical outcomes remain unclear and unstudied.

1.8 Contribution and Outline of the Thesis

In this thesis, we compare different CARA procedures in terms of efficiency and ethics consideration. We recommend the CARA procedures that balance efficiency and ethics better for binary and continuous outcomes. The efficiency is measured by the testing power and ethics is measured by the overall success rate. We then propose a Monte Carlo test for testing the treatment-by-covariate interaction effect that can preserve the type I error rate and maintain power under model misspecification. At the end, we propose a two-stage CARA enrichment design which uses a CARA procedure in the first stage to allocate patients and a Monte Carlo test in the interim analysis to test the treatment-by-covariate interaction effect. The proposed design can preserve the type I error rate, maintain testing power, and have higher overall success rate compared to a standard non-enrichment design under different testing scenarios.

This thesis is structured as follows. In Chapter 2, we compare different CARA procedures for testing treatment-by-covariate interaction effect in terms of efficiency and efficacy. We choose the CARA procedures that balance better between efficiency and ethics. In Chapter 3, we propose a nonparametric simulation-based Monte Carlo test to test the treatment-by-covariate interaction effect. We compare the proposed test to the population model-based tests under different scenarios for binary and continuous outcomes. In Chapter 4, we propose a two-stage enrichment design which uses the selected CARA procedure to allocate patients in the first stage and the Monte Carlo test in the interim analysis. General conclusions and remarks are included in Chapter 5.

Chapter 2: Comparison of Different CARA Procedures

2.1 Binary outcomes in logistic regression models

In this section, we compare the performance of different CARA procedures in testing the interaction effects. First, we consider a binary outcome with two treatment groups and one categorical covariate. Let $X_i = 1$ if a patient's response is a success, and $X_i = 0$ if a patient's response is a failure. Let $p_i = \Pr(X_i = 1|\boldsymbol{Z} = \boldsymbol{z})$ be the probability of success for a given covariate matrix $\boldsymbol{z}$ (which includes the intercept term $z_{i1} = 1$ and group indicators $z_{i2} = T_i$) and $q_i = 1 - p_i$. The logistic regression model is given as

$$\text{logit}(p_i) = \boldsymbol{z}_i\boldsymbol{\beta}, \quad i = 1, \cdots, n, \tag{2.1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of model parameters. Notice that $p = 4$ when there are two covariate strata and $p = 8$ when there are four covariate strata.

Three target allocation rules are considered:

- Rosenberger et al.'s (2001b) target allocation CARA1:

$$\pi_{m+1,1|\boldsymbol{z}_{m+1}} = \frac{p_1(\boldsymbol{z}_{m+1})/q_1(\boldsymbol{z}_{m+1})}{p_1(\boldsymbol{z}_{m+1})/q_1(\boldsymbol{z}_{m+1}) + p_2(\boldsymbol{z}_{m+1})/q_2(\boldsymbol{z}_{m+1})}.$$

- Covariate-adjusted version of Rosenberger et al.'s (2001a) allocation CARA2:

$$\pi_{m+1,1|\boldsymbol{z}_{m+1}} = \frac{\sqrt{p_1(\boldsymbol{z}_{m+1})}}{\sqrt{p_1(\boldsymbol{z}_{m+1})} + \sqrt{p_2(\boldsymbol{z}_{m+1})}}.$$

- Covariate-adjusted version of Rosenberger and Sverdlov's (2008) optimal allocation CARA3:

$$\pi_{m+1,1|\boldsymbol{z}_{m+1}} = \frac{q_2(\boldsymbol{z}_{m+1})\sqrt{p_2(\boldsymbol{z}_{m+1})}}{q_1(\boldsymbol{z}_{m+1})\sqrt{p_1(\boldsymbol{z}_{m+1})} + q_2(\boldsymbol{z}_{m+1})\sqrt{p_2(\boldsymbol{z}_{m+1})}}. \tag{2.2}$$

CARA1 target allocation is proportional to the covariate-adjusted odds ratio. CARA2 minimizes the expected number of treatment failures subject to the fixed asymptotic variance of the test statistic. CARA3 is the optimal allocation which minimizes expected treatment failures subject to the fixed asymptotic variance of the log-odds ratio.

In order to decrease the variability and preserve the randomness of those adaptive procedures which depend on unknown parameters $p_1(\boldsymbol{z})$ and $p_2(\boldsymbol{z})$, doubly-adaptive biased coin design (DBCD) (Hu et al., 2004) and efficient randomized adaptive design (ERADE) (Hu et al., 2009) are used for each target allocation. Covariate-adjusted DBCD allocation rule is defined as:

$$\phi_{m+1,1|z_{m+1}} = \frac{\hat{\pi}_{m+1,1|z_{m+1}}(\frac{\hat{\pi}_{m+1,1|z_{m+1}}}{N_{m+1,1|z_{m+1}}})^\gamma}{\hat{\pi}_{m+1,1|z_{m+1}}(\frac{\hat{\pi}_{m+1,1|z_{m+1}}}{N_{m+1,1|z_{m+1}}})^\gamma + (1-\hat{\pi}_{m+1,1|z_{m+1}})(\frac{1-\hat{\pi}_{m+1,1|z_{m+1}}}{N_{m+1,2|z_{m+1}}})^\gamma},$$

where $N_{n,k|z}$ is the number of patients with given covariate z in group k and γ is the tuning parameter that controls the variability of the procedure. As γ increases,

the procedure becomes less variability and more deterministic. When $\gamma = 0$, this procedure reduces to the sequential MLE CARA procedure.

ERADE allocation rule is defined as:

$$\phi_{m+1,1|z_{m+1}} = \begin{cases} \alpha\hat{\pi}_{m+1,1|z_{m+1}}, & \text{if } N_{m+1,1|z_{m+1}}/m|z_{m+1} > \hat{\pi}_{m+1,1|z_{m+1}} \\ \hat{\pi}_{m+1,1|z_{m+1}}, & \text{if } N_{m+1,1|z_{m+1}}/m|z_{m+1} = \hat{\pi}_{m+1,1|z_{m+1}} \\ 1 - \alpha(1 - \hat{\pi}_{m+1,1|z_{m+1}}) & \text{if } N_{m+1,1|z_{m+1}}/m|z_{m+1} < \hat{\pi}_{m+1,1|z_{m+1}}, \end{cases}$$

where $\alpha \in [0, 1]$. The authors recommend choosing a α between 0.4 and 0.7.

Simulations are programmed in C with $5,000$ replications. Different covariate profiles, such as equally distributed two and four strata and unequally distributed (with ratios of $2 : 8$ and $8 : 2$) two strata are considered in the simulation. Given covariates $\boldsymbol{z}_i$, the response X_i is generated from a Bernoulli distribution with success probability p_i. Under the permuted block randomization (PBR), all cases are equally allocated to two groups with block size of 10. Under CARA procedures, the first 100 cases are allocated using PBR with block size of 10. The likelihood based Wald test and Zhang et al.'s (2007) test in Chapter 1.7 are used to compare the operating characteristics of CARA randomization procedures. The parameter of interest is the treatment-covariate interaction.

For binary outcomes, two treatment groups, and two covariate strata scenarios, the following hypotheses are tested:

$$H_0 : \beta_4 = 0 \text{ versus } H_1 : \beta_4 \neq 0.$$

The parameter values used are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.5$, and $\beta_4 = 0$ under the null hypothesis, and $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.5$, and $\beta_4 = 0.9$ under the alternative hypothesis. The sample size of $1,000$ is chosen so that the equal allocation rule PBR reaches at least 80% power.

For the four covariate strata scenarios, the following hypotheses are tested:

$$H_0 : \beta_6 = \beta_7 = \beta_8 = 0 \text{ versus } H_1 : \text{ Not all three equal to 0.}$$

The parameter values used are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 1, \beta_4 = 0.5$, and $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$ under the null hypothesis, and $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 1, \beta_4 = 0.5, \beta_5 = 0$, and $\beta_6 = \beta_7 = \beta_8 = 0.9$ under the alternative hypothesis. The sample size of $1,500$ is chosen so that PBR achieves at least 80% power.

Due to the higher variability of certain CARA procedures, some scenarios end up having zero success or fail rates during the sequential allocation process. Therefore, Zhang et al.'s (2007) test statistics cannot be calculated and a larger size of the initial allocation is required. In our simulation, under those circumstances, the first 200 or 400 cases are randomized using PBR for two and four covariate strata scenarios respectively.

Table 2.1 shows type I error rates from Wald and Zhang et al's tests for two equally distributed covariate strata and binary outcomes. All allocation procedures preserve type I error rates around the nominal level of 0.05. Simulations are also run under skewed covariate strata scenarios ($2:8$ and $8:2$ distributed) and all type I error rates are found to be around the nominal level of 0.05.

Table 2.2 presents the allocation ratios to the treatment group, overall success rates, and power from Wald and Zhang et al's tests for two equally distributed covariate strata and binary outcomes. Both the sequential MLE and DBCD ($\gamma = 1, 2$)

procedures have higher overall success rates than PBR, while the two DBCDs have smaller variances than the sequential MLE. DBCD procedures also allocate more patients to the better treatment effect group. CARA1 target has higher variability than CARA2 and CARA3. All ERADE procedures allocate more patients to the superior group but have lower power than PBR and DBCD procedures. Under the same effect size, unequally distributed covariates have lower powers across all procedures (Fig.2.1). Similar results are found under the four equally distributed covariate strata situations. When only a small portion of patients benefit more from the treatment, a larger effect size or increased sample size is desired to reach a similar power as in the equally distributed covariate strata scenarios. The parameters used for Table 2.3 are $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.5$, and $\beta_4 = 1.2$. Powers from ERADE procedures are further reduced. Under the sequential MLE and the two DBCD procedures, CARA1 has the least power, highest success rates, and skews the allocation the most; CARA2 has the highest power, lowest success rates, and is closest to the balance design; CARA3 is in the middle. Overall, DBCD($\gamma = 2$) with target CARA3 has a better balance between ethics and efficiency considerations.

Table 2.1: Type I error rates from different randomization designs for two equally distributed covariate strata and binary outcomes

Randomization Procedure	Target Allocation	Wald Error Rate	Zhang Error Rate
PBR		0.048	0.048
DBCD($\gamma = 0$)	CARA1	0.054	0.047
	CARA2	0.054	0.045
	CARA3	0.053	0.045
DBCD($\gamma = 2$)	CARA1	0.050	0.049
	CARA2	0.054	0.053
	CARA3	0.054	0.053
DBCD($\gamma = 1$)	CARA1	0.051	0.050
	CARA2	0.055	0.055
	CARA3	0.052	0.052
ERADE($\alpha = 0.5$)	CARA1	0.046	0.045
	CARA2	0.049	0.048
	CARA3	0.046	0.045
ERADE($\alpha = 0.6$)	CARA1	0.049	0.047
	CARA2	0.050	0.050
	CARA3	0.049	0.049
ERADE($\alpha = 0.7$)	CARA1	0.049	0.046
	CARA2	0.053	0.053
	CARA3	0.049	0.049

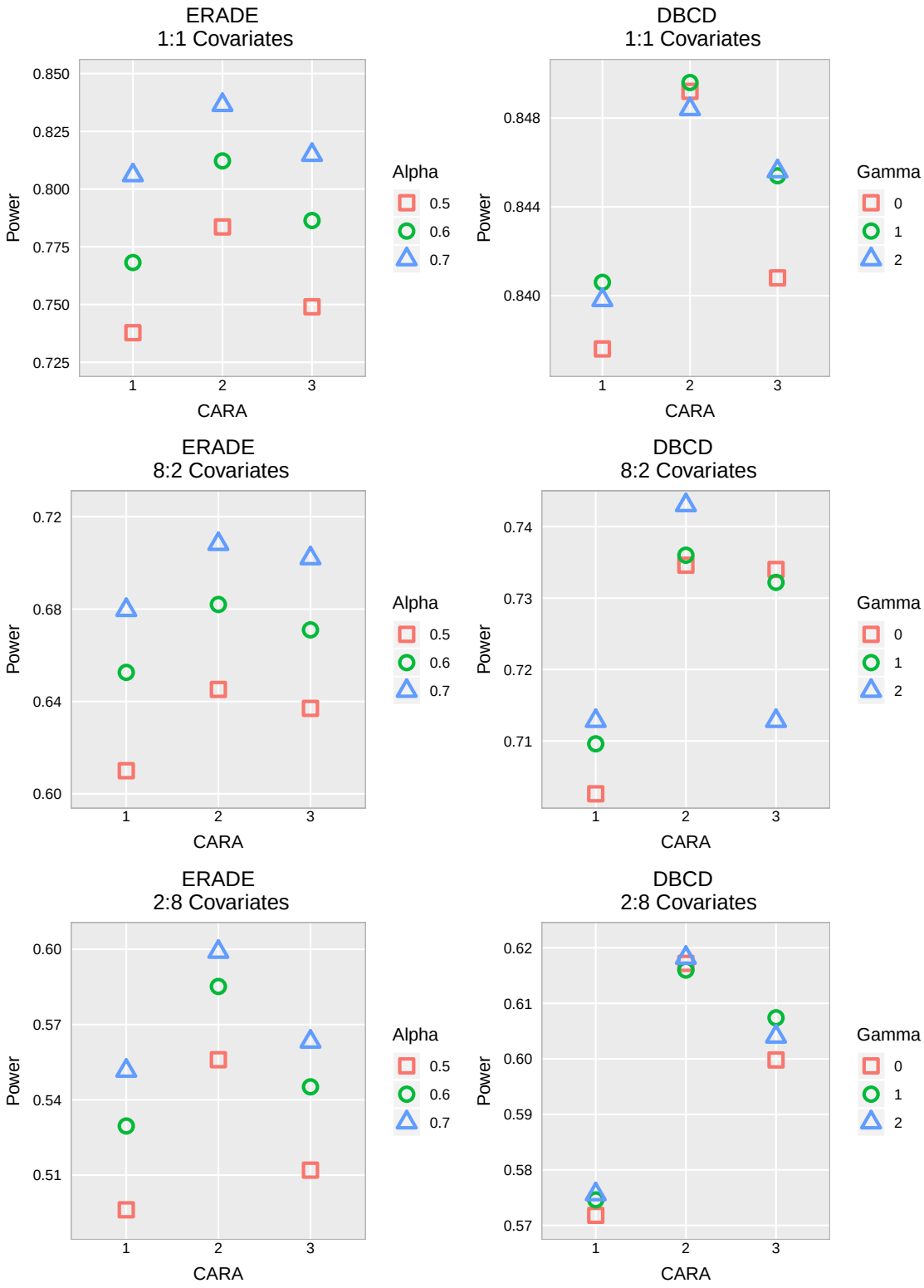

Figure 2.1: Power from different CARA procedures, 5000 runs, n = 1000

Table 2.2: Power, allocation ratios, and overall success rates from different randomization designs for two equally distributed covariate strata and binary outcomes

Randomization	Target Allocation	Wald Power	Zhang Power	SR[1](var*N)	AR[2](var*N)
PBR		0.861	0.860	0.712(0.206)	0.500(0.000)
DBCD($\gamma = 0$)	CARA1	0.838	0.842	0.725(0.220)	0.594(1.766)
	CARA2	0.849	0.849	0.713(0.207)	0.510(0.268)
	CARA3	0.840	0.838	0.721(0.221)	0.569(0.836)
DBCD($\gamma = 2$)	CARA1	0.840	0.833	0.726(0.217)	0.605(1.489)
	CARA2	0.848	0.849	0.713(0.205)	0.511(0.085)
	CARA3	0.846	0.839	0.722(0.219)	0.576(0.591)
DBCD($\gamma = 1$)	CARA1	0.841	0.841	0.726(0.217)	0.603(1.592)
	CARA2	0.850	0.850	0.713(0.205)	0.511(0.123)
	CARA3	0.845	0.840	0.722(0.219)	0.576(0.655)
ERADE($\alpha = 0.5$)	CARA1	0.738	0.719	0.734(0.203)	0.772(0.592)
	CARA2	0.784	0.783	0.728(0.202)	0.730(0.183)
	CARA3	0.749	0.740	0.732(0.206)	0.759(0.316)
ERADE($\alpha = 0.6$)	CARA1	0.768	0.753	0.732(0.206)	0.737(0.760)
	CARA2	0.812	0.811	0.725(0.202)	0.686(0.207)
	CARA3	0.786	0.775	0.730(0.208)	0.721(0.398)
ERADE($\alpha = 0.7$)	CARA1	0.806	0.791	0.730(0.209)	0.701(0.966)
	CARA2	0.836	0.836	0.722(0.203)	0.642(0.232)
	CARA3	0.815	0.806	0.728(0.211)	0.683(0.491)

[1] success rate
[2] allocation ratio

2.2 Continuous outcomes in linear regression models

We then consider continuous normal outcomes with two treatment groups and one categorical covariate. Suppose responses follow a linear regression model with homoscedastic variance

$$X_i = \boldsymbol{Z}_i\boldsymbol{\beta} + \epsilon, \tag{2.3}$$

where $\epsilon \sim N(0, \sigma^2)$, X_i is a $n \times 1$ vector of responses, $\boldsymbol{Z}_i$ is a $n \times p$ covariate matrix (including the intercept $z_{i1} = 1$ and group indicators $z_{i2} = T_i$), and $\boldsymbol{\beta}$ is a $p \times 1$ vector of model parameters. Assuming that larger responses are desirable, the target allocation which is based on Zhang et al.'s (2007) CARA design is calculated as:

$$\pi_{m+1,1|z_{m+1}} = \Phi\left(\frac{\boldsymbol{Z}_{m+1}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{Z}_{m+1}\hat{\boldsymbol{\beta}}_2}{G}\right), \tag{2.4}$$

where $\boldsymbol{\beta}_k$ are the parameter estimators for the kth group ($k = 1, 2$) and G is a tuning parameter. The smaller the T value, the more the allocation is skewed. We compare $G = 2$ and 6 in our simulations.

For continuous outcomes, two treatment groups, and two covariate strata scenarios, the following hypotheses are tested:

$$H_0 : \beta_4 = 0 \text{ versus } H_1 : \beta_4 \neq 0.$$

The parameter values used are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.5$, and $\beta_4 = 0$ under the null hypothesis, and $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.5$, and $\beta_4 = 0.6$ under the alternative hypothesis. For the four covariate strata scenarios, the following hypotheses are

Table 2.3: Power, allocation ratios, and overall success rates from different randomization designs for two unequally (2 : 8) distributed covariate strata and binary outcomes

Randomization	Target Allocation	Wald Power	SR[1](var*N)	AR[2](var*N)
PBR		0.830	0.661(0.225)	0.500(0.000)
DBCD($\gamma = 0$)	CARA1	0.759	0.670(0.237)	0.551(1.779)
	CARA2	0.822	0.662(0.227)	0.505(0.284)
	CARA3	0.809	0.668(0.232)	0.530(0.655)
DBCD($\gamma = 2$)	CARA1	0.761	0.671(0.233)	0.556(1.456)
	CARA2	0.826	0.662(0.225)	0.505(0.099)
	CARA3	0.816	0.668(0.231)	0.541(0.438)
DBCD($\gamma = 1$)	CARA1	0.764	0.671(0.233)	0.555(1.594)
	CARA2	0.823	0.662(0.225)	0.505(0.136)
	CARA3	0.812	0.668(0.230)	0.540(0.491)
ERADE($\alpha = 0.5$)	CARA1	0.682	0.673(0.225)	0.749(0.615)
	CARA2	0.762	0.669(0.230)	0.728(0.192)
	CARA3	0.703	0.672(0.226)	0.743(0.285)
ERADE($\alpha = 0.6$)	CARA1	0.722	0.672(0.226)	0.708(0.779)
	CARA2	0.784	0.668(0.224)	0.683(0.217)
	CARA3	0.742	0.671(0.227)	0.702(0.345)
ERADE($\alpha = 0.7$)	CARA1	0.751	0.671(0.226)	0.668(0.973)
	CARA2	0.807	0.666(0.224)	0.639(0.242)
	CARA3	0.758	0.670(0.228)	0.661(0.420)

[1] success rate
[2] allocation ratio

tested:

$$H_0 : \beta_6 = \beta_7 = \beta_8 = 0 \text{ versus } H_1 : \text{ Not all three equal to 0.}$$

The parameter values used are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 1, \beta_4 = 0.5$, and $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$ under the null hypothesis, and $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 1, \beta_4 = 0.5, \beta_5 = 0$, and $\beta_6 = \beta_7 = \beta_8 = 0.8$ under the alternative hypothesis. Different covariate profiles, such as equally distributed two and four strata and unequally distributed (with ratios of $3:7$ and $7:3$) two strata are considered in the simulation. Given covariates $\boldsymbol{z}_i$, the response X_i is generated from the equation 2.3. The residuals are generated from a $N(0,1)$ distribution. Ten percent of outliers are randomly generated from a $N(-5,4)$ distribution and used in the outlier scenarios.

Under Wald tests, with or without outliers, all DBCD and ERADE procedures preserve type I error rates when there are two covariate strata. Conservative type I error rates are observed when Wald tests are used to test the interaction effect for four covariate strata. When there are outliers in four strata scenarios, tuning parameter of 2 produces Inflated type I error rates in DBCD procedures as well. Zhang et al's tests have inflated type I error rates under all scenarios when outliers are presented. Table 2.4 shows type I error rates from Wald and Zhang et al's tests for four equally distributed covariate strata and 10% outliers. Power and allocation ratios for four equally distributed covariates are presented in Table 2.5. Zhang et al's tests have higher power from all allocations rules. Sequential MLE and the two DBCD procedures allocate more patients to the superior group. All ERADE procedures allocate more patients to the superior group but have lower powers. DBCD($\gamma = 2$) targeting CARA($T = 6$) has the smallest variability and highest power. Outliers significantly reduce testing powers for both Wald and Zhang et al's tests under all allocation rules. Procedures with smaller tuning parameters allocate more patients to

the treatment group, but also have higher variabilities. Overall, DBCD($\gamma = 2$) with a tuning parameter of 6 performs the best among all the procedures compared. Table 2.6 presents power and allocation ratios to the treatment group from Wald and Zhang et al's tests for four equally distributed covariate strata with 10% outliers. As shown in figure 2.2, under the same effect size, for all DBCD procedures, testing powers are significantly higher when there are no outliers and covariates are equally distributed.

Table 2.4: Type I error rates from different randomization designs for four equally distributed covariate strata and continuous outcomes with 10% outliers

Randomization Procedure	Target Allocation	Wald Error Rate	Zhang Error Rate
PBR		0.039	0.544
DBCD($\gamma = 0$)	CARA(T=6)	0.043	0.541
	CARA(T=2)	0.060	0.540
DBCD($\gamma = 2$)	CARA(T=6)	0.045	0.536
	CARA(T=2)	0.074	0.556
DBCD($\gamma = 1$)	CARA(T=6)	0.043	0.537
	CARA(T=2)	0.070	0.547
ERADE($\alpha = 0.5$)	CARA(T=6)	0.046	0.533
	CARA(T=2)	0.055	0.532
ERADE($\alpha = 0.6$)	CARA(T=6)	0.046	0.530
	CARA(T=2)	0.056	0.543
ERADE($\alpha = 0.7$)	CARA(T=6)	0.046	0.537
	CARA(T=2)	0.055	0.542

2.3 Conclusion

We use numerical studies to compare the performance of different CARA randomization procedures for testing the treatment-by-covariate interaction effect. All procedures preserve type I error rates when there are no model misspecification. Inflated type I error rates and reduced testing power are observed when there are outliers in the linear regression models. ERADE procedures allocate more patients to the superior group, therefore have higher success rates than PBR and DBCD procedures. However, the powers from ERADE procedures are generally lower than PBR and DBCD procedures as well.

For binary outcomes, DBCD ($\gamma = 2$) targeting CARA3 is chosen since it has a better balance between the power and the overall success rate. For continuous outcomes, DBCD ($\gamma = 2$) targeting the CARA procedure with a tuning parameter of

Table 2.5: Power and allocation ratios from different randomization designs for four equally distributed covariate strata and continuous outcomes

Randomization	Target Allocation	Wald Power	Zhang Power	AR[1](var*N)
PBR		0.684	0.835	0.500(0.000)
DBCD($\gamma = 0$)	CARA(T=6)	0.667	0.824	0.530(0.198)
	CARA(T=2)	0.654	0.809	0.587(0.296)
DBCD($\gamma = 2$)	CARA(T=6)	0.667	0.832	0.539(0.071)
	CARA(T=2)	0.648	0.817	0.614(0.221)
DBCD($\gamma = 1$)	CARA(T=6)	0.670	0.827	0.537(0.100)
	CARA(T=2)	0.646	0.814	0.608(0.243)
ERADE($\alpha = 0.5$)	CARA(T=6)	0.551	0.756	0.702(0.138)
	CARA(T=2)	0.532	0.744	0.731(0.152)
ERADE($\alpha = 0.6$)	CARA(T=6)	0.587	0.779	0.668(0.156)
	CARA(T=2)	0.564	0.769	0.702(0.182)
ERADE($\alpha = 0.7$)	CARA(T=6)	0.614	0.799	0.633(0.170)
	CARA(T=2)	0.601	0.783	0.673(0.212)

[1] allocation ratio

Table 2.6: Power and allocation ratios from different randomization designs for four equally distributed covariate strata and continuous outcomes with 10% outliers

Randomization Procedure	Target Allocation	Wald Power	Allocation Ratio (var*N)
PBR		0.256	0.500(0.000)
DBCD($\gamma = 0$)	CARA(T=6)	0.255	0.529(0.237)
	CARA(T=2)	0.254	0.585(0.640)
DBCD($\gamma = 2$)	CARA(T=6)	0.251	0.539(0.126)
	CARA(T=2)	0.273	0.612(0.747)
DBCD($\gamma = 1$)	CARA(T=6)	0.253	0.537(0.153)
	CARA(T=2)	0.271	0.605(0.729)
ERADE($\alpha = 0.5$)	CARA(T=6)	0.214	0.702(0.148)
	CARA(T=2)	0.222	0.730(0.233)
ERADE($\alpha = 0.6$)	CARA(T=6)	0.224	0.668(0.170)
	CARA(T=2)	0.242	0.702(0.299)
ERADE($\alpha = 0.7$)	CARA(T=6)	0.244	0.633(0.190)
	CARA(T=2)	0.251	0.673(0.376)

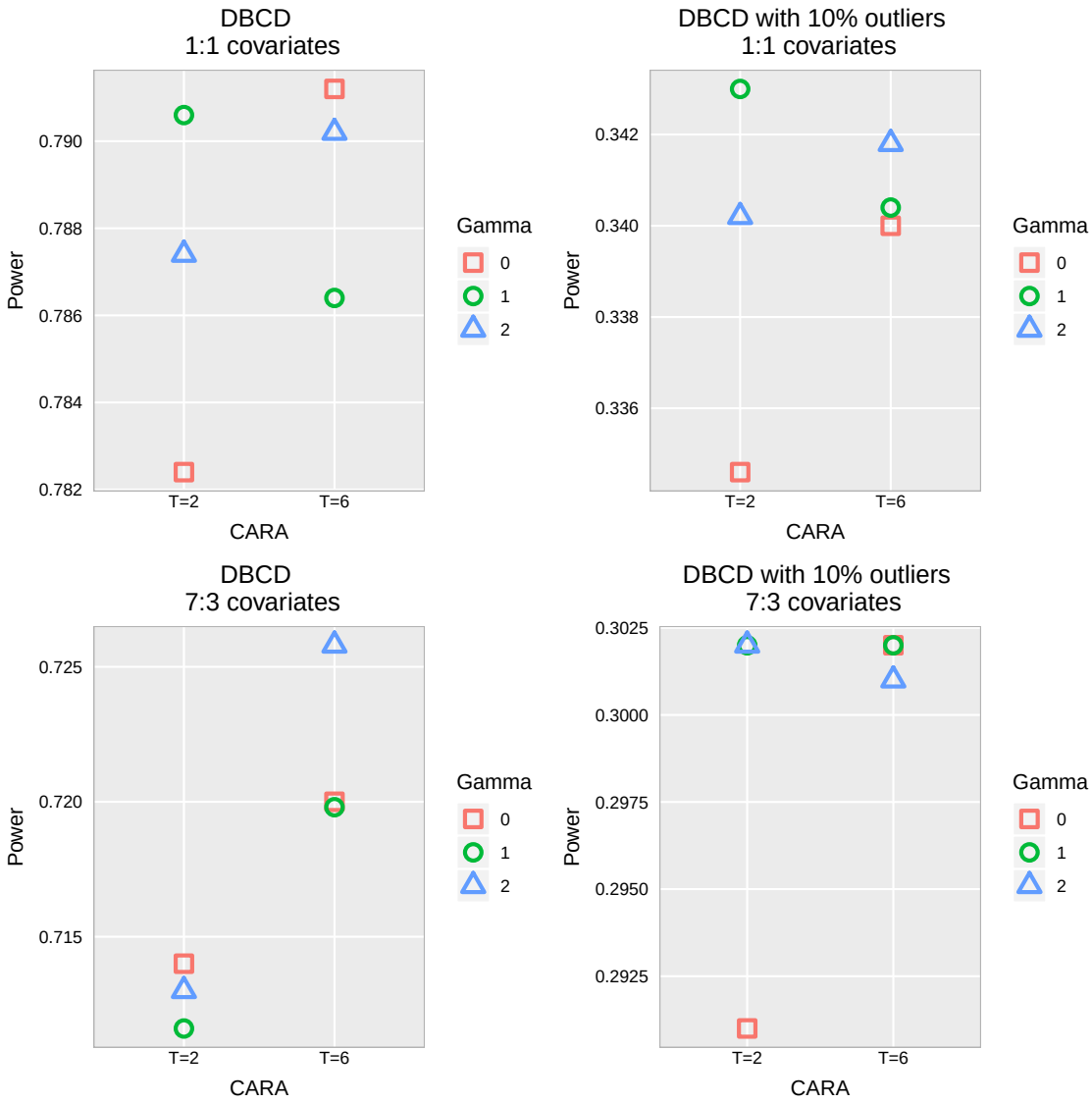

Figure 2.2: Power from different DBCD procedures, 5000 runs, n = 400

6 is chosen since it has the higher power with moderately low allocation variability.

Chapter 3: Monte Carlo Tests of Interaction Effect

Based on the findings from previous chapters, population-based tests encounter inflated type I error rates and reduced power under model misspecification. Randomization tests preserve the size and power under model misspecification for testing treatment effects (Parhat et al., 2014). In this chapter, we try to find a similar application to test the interaction effect.

When testing the interaction effect, under the null hypothesis, there still might be overall treatment effects. In this case, we cannot assume that the treatment allocation is unrelated to the outcome, therefore, permuting the treatment allocation to run a randomization test is not applicable. Another commonly used simulation-based test is the permutation test. It assumes exchangeability and can be used for both randomized and non-randomized data sets. Under a permutation test, the probability of each permuted sequence is the same. Since we generally cannot assume the exchangeability condition under a CARA procedure, permutation tests are not applicable as well.

Here, we consider a generalized linear model (GLM) setting for two groups. Suppose for a given covariate $\boldsymbol{Z}$, the response X_k of the treatment $k = 1, 2$ has a distribution in the exponential family under the GLM:

$$f_k(X_k|\boldsymbol{Z}, \boldsymbol{\theta}_k) = \exp\{(X_k\mu_k - \alpha_k(\mu_k))/\psi_k + b_k(X_k, \psi_k)\},$$

with the link function $\mu_k = h_k(\boldsymbol{Z}\boldsymbol{\theta}_k')$, where $\theta_k = (\theta_{k1}, \cdots, \theta_{kd}),\ k = 1, 2$. Assume the scale parameter ψ_k is fixed. Under this model, $E(X_k|Z) = \alpha'(\mu_k),\ \mathrm{Var}(X_k|Z) =$

$\alpha''(\mu_k)\psi_k$. The first and second derivatives for log likelihood function are:

$$\frac{\partial log f_k(X_k|\boldsymbol{Z},\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k} = \frac{1}{\psi_k}(X_k - a_k'(\mu_k))h_k'(\boldsymbol{Z}\boldsymbol{\theta}_k')\boldsymbol{Z}$$

and

$$\frac{\partial^2 log f_k(X_k|\boldsymbol{Z},\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k^2} = \frac{1}{\psi_k}\{-a_k''(\mu_k))[h_k'(\boldsymbol{Z}\boldsymbol{\theta}_k')]^2 + [X_k - a_k'(\mu_k)]h_k''(\boldsymbol{Z}\boldsymbol{\theta}_k')\}\boldsymbol{Z}'\boldsymbol{Z}.$$

After using a CARA procedure to allocate all n patients to the two treatment groups and collecting the responses, a generalized linear regression model is fit. The score test statistic for testing the interaction effect based on the observed dataset is calculated as the observed test statistic. The score test only involves the restricted maximum likelihood estimation and the test statistic is calculated as:

$$S_n = [L_1(\theta, X_n)'(-L_2(\theta, X_n)^{-1}L_1(\theta, X_n)]_{\theta=\tilde{\theta}},$$

where $L_1(\theta, X_n)$ is the first derivative of the log likelihood function and $L_2(\theta, X_n)$ is the second derivative of the log likelihood function.

While keeping the treatment allocation and the covariate strata status fixed, we regenerate L_s sequences of the interaction terms using Monte Carlo simulation and the score test statistics are calculated for each set of permutations. The two-sided p-value is calculated as

$$\hat{p} = \frac{\sum_{l=1}^{L_s} I(|S_l| \geq |S_{obs}|)}{L_s}.$$

The value of L at $2,500$ will bound the mean squared error of the p-value at 0.0001 . In order to estimate very small p-values accurately, Rosenberger and Lachin (2015)

suggest $20,000$ sequences. Galbete and Rosenberger (2016) demonstrate that $15,000$ sequences produce tests that are almost identical to exact tests. This test has no validity under a randomization or permutation world, however, it will preserve the type I error rate. We refer to this test as a "Monte Carlo test".

3.1 Binary responses

Simulations are run to compare the performance of population model-based approaches and the proposed Monte Carlo test for both binary and continuous outcome variables. First we consider a binary outcome with two treatment groups in a logistic regression model (equation 2.1).

Let the group by covariate interaction $Y_i = T_i Z_i$. A total of $15,000$ permutations of Y_i are generated with fixed covariates, responses, and group allocations. Score test statistics are calculated from the observed and permuted data sets and the corresponding p-values are then calculated. Based on the findings from Chapter 2, DBCD ($\gamma = 2$) with target CARA3 (equation 2.2) is used in the simulation. The parameter values used in the two covariate strata scenarios are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.5$, and β_4 ranging from 0 to 0.7. The sample size of $1,000$ is chosen so that the Wald test achieves at least 80% power when $\beta_4 = 1.2$ and there are more patients from the worse responsive stratum ($8:2$ covariate ratio). Figure 3.1 presents the results from simulations. All three tests have similar rejection rates. Zhang et al's tests under both CARA and PBR have slightly lower powers when only a small portion of patients benefit from the treatment under the CARA allocation.

The parameter values used in the four covariate strata scenarios are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 1, \beta_4 = 0.5, \beta_5 = 0$, and $\beta_6 - \beta_8$ ranging from 0 to 1. The sample size of $1,500$ is chosen so that the Wald test achieves at least 80% power when $\beta_6 = \beta_7 = \beta_8 = 1$.

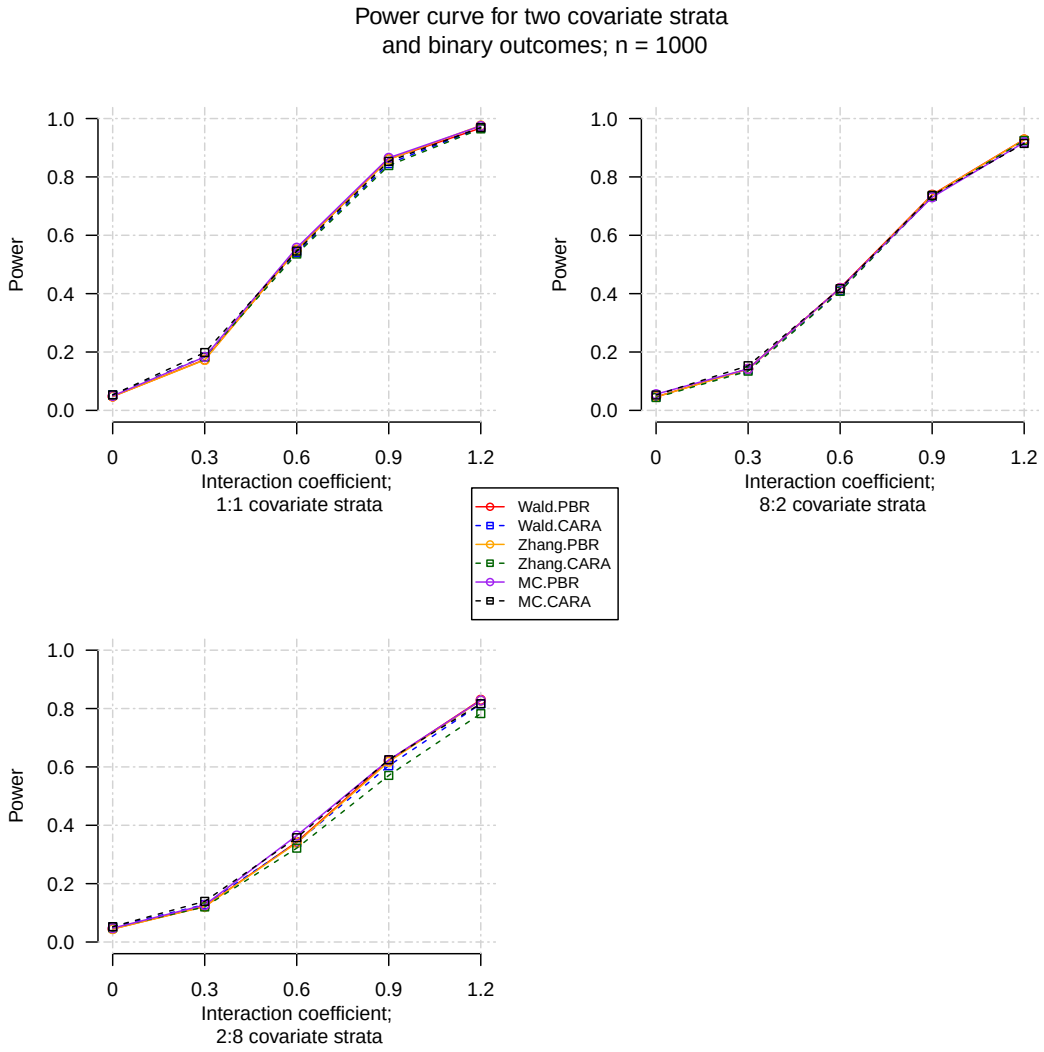

Figure 3.1: Two covariate strata, two treatment groups, and binary outcomes

Zhang et al's test also shows slightly lower power under the CARA allocation (Figure 3.2), while Wald and Monte Carlo tests have similar power under both CARA and PBR allocations.

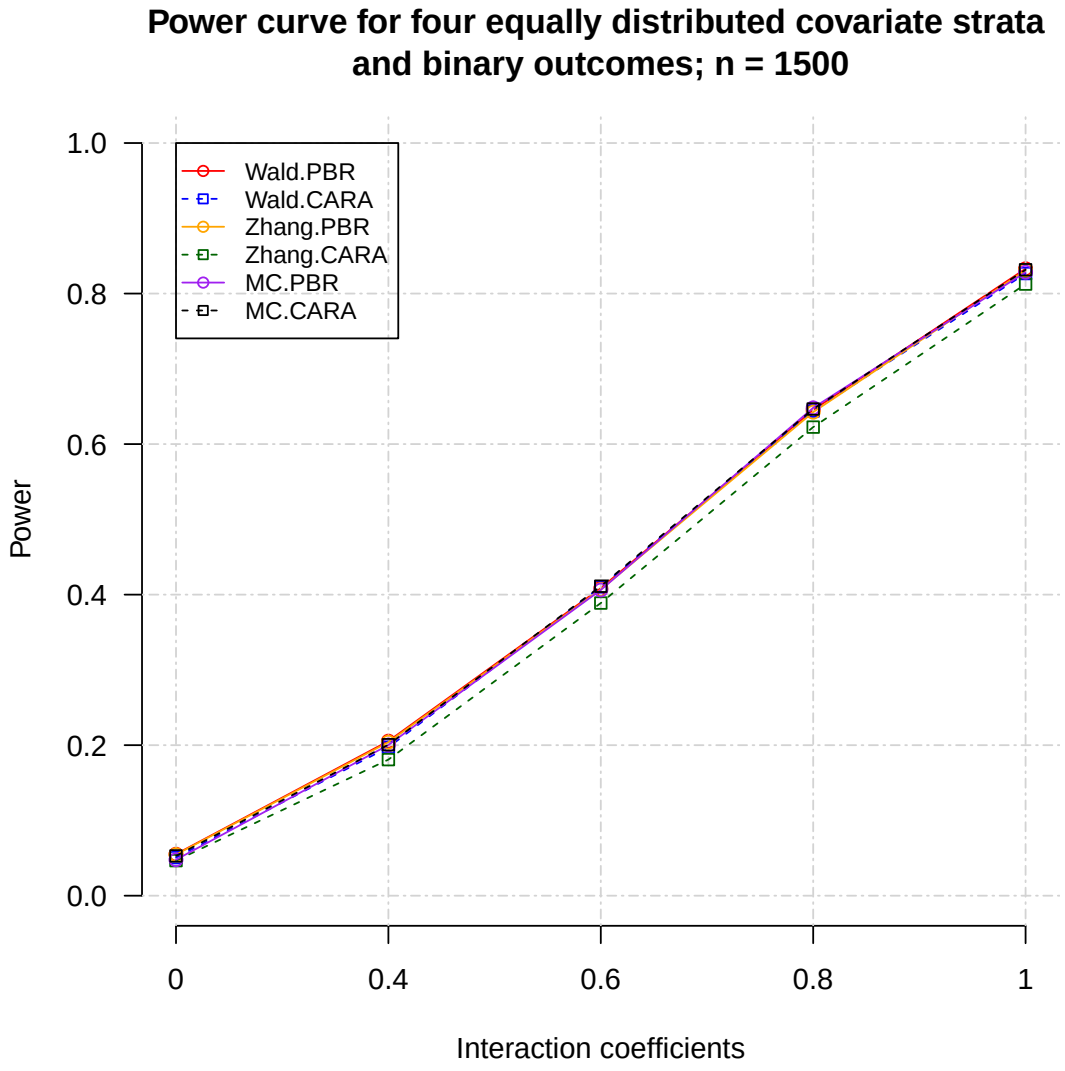

Figure 3.2: Four covariate strata, two treatment groups, and binary outcomes

3.2 Continuous responses

We now consider a continuous outcome with two treatment groups in a linear regression model (equation 2.3). Based on the findings from Chapter 2, DBCD ($\gamma = 2$) with target CARA ($T = 6$) (equation 2.4) is used in the simulation. The parameter values used in the two covariate strata scenarios are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.5$, and

β_4 ranging from 0 to 1.2. The residuals are generated from a $N(0,1)$ distribution. Ten percent of outliers are randomly generated from a $N(-5,4)$ distribution and used in the outlier scenarios. The sample sizes are chosen so that the Wald test achieves at least 80% power. Figure 3.3 presents the results of simulations. All three tests have similar rejection rates when no outliers are presented. To reach the same power level, unequally distributed covariates require larger sample sizes. Wald tests have slightly lower power under both PBR and CARA allocations. Zhang et al's tests have inflated type I error rates when outliers are presented. Although both Wald and Monte Carlo tests preserve type I error rates around the nominal level, Monte Carlo tests consistently have higher power than Wald tests.

The parameter values used in the four covariate strata scenarios are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 1, \beta_4 = 0.5, \beta_5 = 0$, and $\beta_6 - \beta_8$ ranging from 0 to 1.5. All three tests preserve type I error rates under the no outlier scenario, and Wald test has lower power than the other two tests. Zhang et al's test has inflated type 1 error rates and Wald test has lower power when outliers are presented (figure 3.4).

3.3 Conclusion

Population model-based and the proposed Monte Carlo tests perform equally well when there are two strata and no misspecified data. However, when there are multiple responsive covariate strata or outliers, inflated type I error rates and reduced power are observed under the population model-based tests. Although the score test statistics are used in calculating the p-values, the procedure itself is nonparametric since it is based on the randomization distribution induced b the particular sequence.

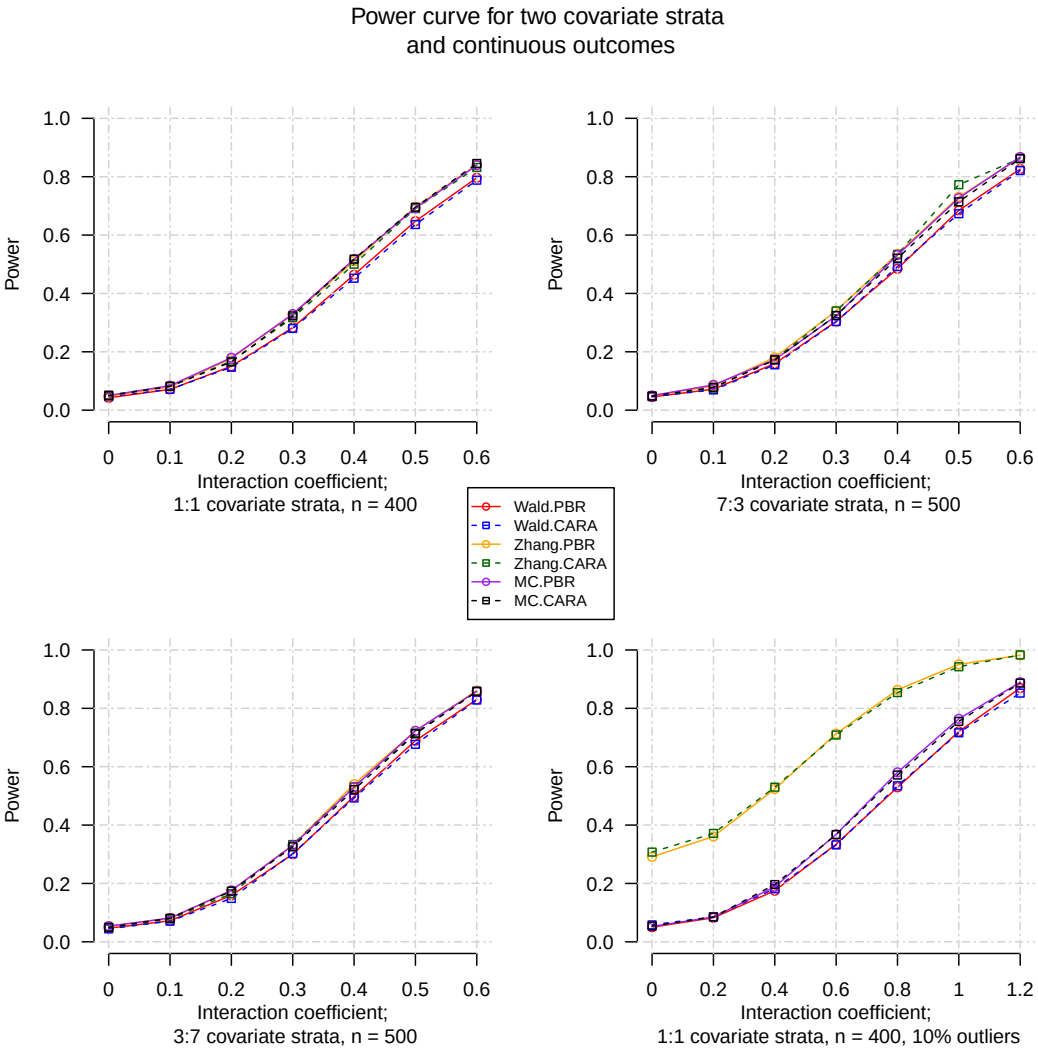

Figure 3.3: Two covariate strata, two treatment groups, and continuous outcomes

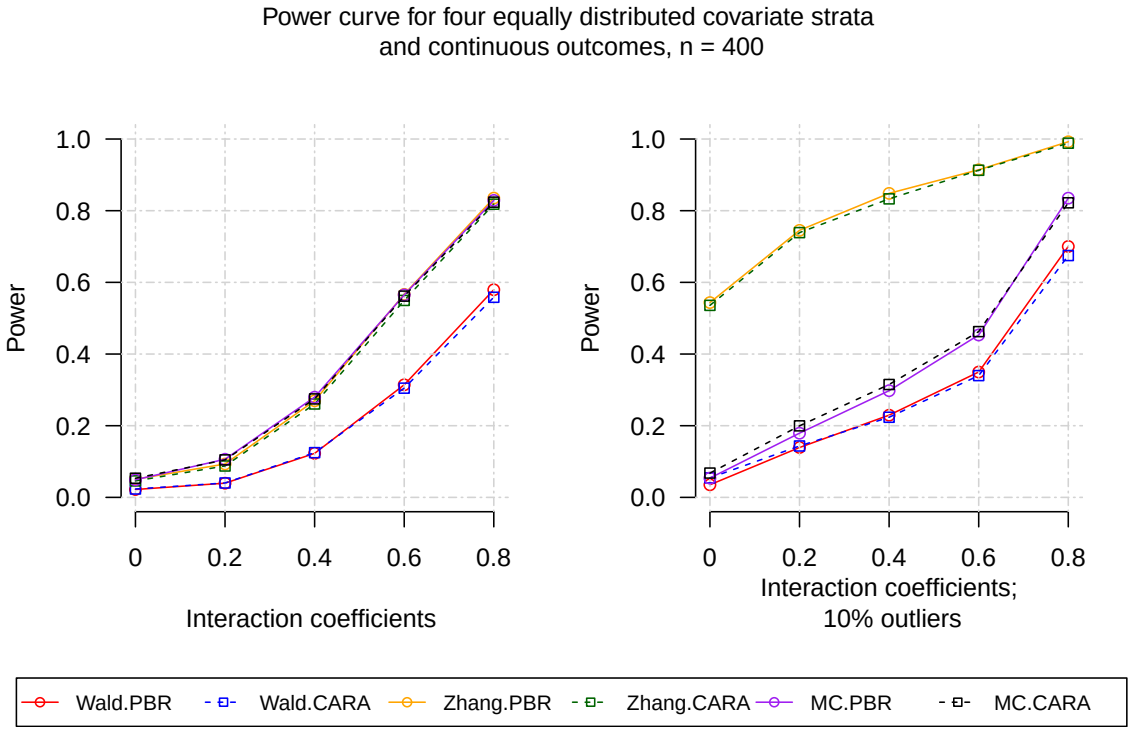

Figure 3.4: Four covariate strata, two treatment groups, and continuous outcomes

Chapter 4: A two-stage Enrichment Design using Monte Carlo Tests

Despite of the existence of treatment-by-covariate interaction effect in many situations, clinical trials are often designed under the assumption of no such effect. Ayanlowo and Redden's (2008) two-stage adaptive design examines the interaction effect and a trial moves to the second stage only if a significant interaction effect is detected at the interim analysis. The second stage is non-enriched, but stratified based on covariate strata. Simon and Simon's (2013) design selects the responsive patients in the second stage. If some but not all sub-groups respond to the treatment, it is an enrichment design. If all cases benefit from the treatment, it becomes a non-enrichment design. Freidlin et al. (2010) summarized the characteristics for randomized clinical trials with biomarkers. They noted that one of the main limitations of the classical enrichment design is that the biomarker might not be able to identify the subgroup of patients who benefit with reasonable accuracy. In this chapter, we propose a two-stage enrichment design which uses the Monte Carlo test to evaluate the interaction effect in the interim analysis. Similar to Simon and Simon's (2013) design and Diao et al.'s (2018) design, our design does not adaptively adjust the total sample size after the first stage. The enrichment in the second stage is expected to increase power for hypothesis testing using either data from the second stage alone or combined data from both stages.

We still consider a generalized linear model setting for two groups. Suppose for a given covariate $\boldsymbol{Z}$, the response X_k of the treatment $k = 1, 2$ has a distribution in

the exponential family under the model:

$$f_k(X_k|\boldsymbol{Z},\boldsymbol{\theta}_k) = \exp\{(X_k\mu_k - \alpha_k(\mu_k))/\psi_k + b_k(X_k,\psi_k)\},$$

with the link function $\mu_k = h_k(\boldsymbol{Z}\boldsymbol{\theta}_k')$, where $\theta_k = (\theta_{k1},\cdots,\theta_{kd}),\ k = 1,2$. For binary outcomes, let p_1 be the success rate for the treatment group and p_2 the success rate for the control group. The treatment effect size Δ is defined by the relative risk p_1/p_2. Similarly, in each stratum g, let p_{1g} be the success rate for the treatment group and p_{2g} the success rate for the control group. The effect size of the gth subset is $\Delta_g = p_{1g}/p_{2g}$. For continuous outcomes, let μ_1 be the mean response of the primary efficacy outcome for the treatment group and μ_2 the mean response for the control group. Assume that the response variable in the treatment and the control groups has an equal variance denoted by σ^2. The treatment effect size Δ is defined by $\Delta = (\mu_1 - \mu_2)$. Similarly, in each stratum g, let μ_{1g} be the mean response for the treatment group, μ_{2g} the mean response for the control group, and σ_g^2 be the common variance. The effect size of the gth subset is $\Delta_g = (\mu_{1g} - \mu_{2g})$.

The proposed two-stage adaptive enrichment design uses the Monte Carlo test described in Chapter 3 to test the interaction effect at the interim analysis. A pre-defined alpha level is used as the threshold to decide whether a subgroup will be identified and recruited in the second stage. If a significant interaction effect is found, a regression model will be fitted and the stratum with the largest treatment effect will be chosen as the best stratum. The trial will continue to the second stage with patients from the best stratum only. The response-adaptive randomization (RA) version of DBCD ($\gamma = 2$) with CARA3 will be used to allocate the rest of the patients. If the p-value from the interim analysis is above the threshold, the trial continues with all patients. The primary aim is to test the treatment effect between treatment groups.

4.1 Binary responses

For a binary outcome with two treatment groups in a logistic regression model (equation 2.1), simulations are run to compare the proposed design with the traditional non-enrichment design and Simon and Simon's enrichment design (Simon and Simon, 2013). Wald test and randomization test are used in the final analysis for the proposed design. The parameters used in the two covariate strata scenarios are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.5$, and β_4 ranging from 0 to 1.1. The sample size of $1,000$ is chosen so that the non-enrichment design achieves at least 80% power when $\beta_4 = 1.1$.

Table 4.1 shows the results from two equally distributed covariate strata. All four designs preserve type I error rates. Under the null hypothesis ($\beta_4 = 0$) or effect sizes are small ($\beta_4 = 0.3$), Simon and Simon's enrichment designs could end up with smaller sample sizes since neither covariate strata meets the inclusion criterion (equation 1.12), therefore it could be less powerful than the standard non-enrichment design. As the effect size gets larger ($\beta_4 = 0.5, 0.7$), Simon and Simon's enrichment design becomes more powerful than the non-enrichment design. The overall success rates in Simon and Simon's enrichment designs are always higher than non-enrichment designs due to the biased sampling in the second stage. The two Monte Carlo enrichment designs consistently have higher powers than non-enrichment and Simon and Simon's enrichment designs under different effect sizes. Monte Carlo enrichment designs with non-enriched second stages have lower overall success rates than Simon and Simon's enrichment designs, but those with enriched second stages have higher overall success rates. When there are only a small portion of patients who would benefit from the treatment (Table 4.2), the two Monte Carlo enrichment designs with enriched second stages ($\beta_4 = 0.7$) have much higher power than non-enrichment and Simon and Simon's enrichment designs. The two Monte Carlo enrichment designs have

similar powers under all scenarios. A larger threshold increases the probability of a enriched second stage. The variances of the success rates under Simon and Simon's enrichment designs are consistently higher than the other designs.

The parameter values used in the four covariate strata scenarios are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 1, \beta_4 = 0.5, \beta_5 = 0$, and $\beta_6 - \beta_8$ ranging from 0 to 0.7. The sample size of $1,500$ is chosen so that the non-enrichment design achieves at least 80% power when $\beta_6 = \beta_7 = \beta_8 = 0.5$. When there are four equally distributed strata and three of the four strata have higher success rates in the treatment group (Table 4.3), non-enrichment designs have higher power than Simon and Simon's enrichment designs when the effect sizes are small. As the effect sizes increase, Simon and Simon's enrichment designs become more powerful than non-enrichment designs. Monte Carlo enrichment designs have higher power than non-enrichment designs, however the overall success rates are lower since the third covariate group is chosen and enriched based on the interim analysis results. Although as noted in Diao et al. (2018), using the best group only in the final analysis leads to inflated type I error rates, type I error rate inflation is not observed in our simulations. Since the power of a test depends on both the sample size in the final analysis and the effect size, for the selected parameters, under the equal allocation, using all data is more powerful than using the second stage data. Using second stage data only is more powerful than using all data when there is only a small portion of the patients who respond better to the treatment (Fig 4.1).

Table 4.1: Power and overall success rates from different designs for two equally distributed covariate strata and binary outcomes

Beta	Design	Threshold	Rejection Rate	SR[1](var*N)
(0.5, 0, 0.5, 0)	NED[2]		0.051	0.677(0.222)
	SED[3]		0.041	0.677(0.578)
	MCED1[4]	0.2 & 0.4	0.057	0.677(0.223)
	MCED2[5]	0.2 & 0.4	0.057	0.677(0.222)
(0.5, 0, 0.5, 0.3)	NED		0.159	0.690(0.215)
	SED		0.151	0.705(0.628)
	MCED1	0.2	0.215	0.692(0.221)
	MCED2	0.2	0.217	0.692(0.216)
	MCED1	0.4	0.374	0.726(0.203)
	MCED2	0.4	0.371	0.726(0.204)
(0.5, 0, 0.5, 0.5)	NED		0.327	0.698(0.210)
	SED		0.401	0.721(0.575)
	MCED1	0.2	0.491	0.702(0.218)
	MCED2	0.2	0.485	0.702(0.216)
	MCED1	0.4	0.750	0.742(0.198)
	MCED2	0.4	0.757	0.742(0.201)
(0.5, 0, 0.5, 0.7)	NED		0.517	0.705(0.207)
	SED		0.715	0.732(0.555)
	MCED1	0.2 & 0.4	0.947	0.757(0.193)
	MCED2	0.2 & 0.4	0.956	0.757(0.197)

[1] success rate
[2] non-enrichment design
[3] Simon and Simon's enrichment design
[4] Monte Carlo enrichment design using Wald test
[5] Monte Carlo enrichment design using randomization test

4.2 Continuous responses

For a continuous outcome with two treatment groups in a linear regression model (equation 2.3), simulations are run to compare non-enrichment designs, Monte Carlo enrichment designs, and Simon and Simon's enrichment design. Since there are no explicit test statistics given by Simon and Simon (2013) for continuous outcomes in a sequential enrollment scenario, the group sequential analysis method with block size

Table 4.2: Power and overall success rates from different designs for two unequally (2 : 8) distributed covariate strata and binary outcomes

Beta	Design	Threshold	Rejection Rate	SR[1](var*N)
(0.5, 0, 0.5, 0)	NED[2]		0.052	0.645(0.223)
	SED[3]		0.048	0.649(0.585)
	MCED1[4]	0.2 & 0.4	0.050	0.644(0.231)
	MCED2[5]	0.2 & 0.4	0.052	0.644(0.231)
(0.5, 0, 0.5, 0.3)	NED		0.069	0.650(0.231)
	SED		0.062	0.668(0.983)
	MCED1	0.2	0.087	0.650(0.231)
	MCED2	0.2	0.082	0.650(0.230)
	MCED1	0.4	0.283	0.705(0.207)
	MCED2	0.4	0.276	0.705(0.209)
(0.5, 0, 0.5, 0.5)	NED		0.091	0.653(0.230)
	SED		0.156	0.681(1.129)
	MCED1	0.2 & 0.4	0.141	0.654(0.232)
	MCED2	0.2 & 0.4	0.135	0.654(0.231)
(0.5, 0, 0.5, 0.7)	NED		0.121	0.656(0.229)
	SED		0.332	0.695(1.217)
	MCED1	0.2 & 0.4	0.855	0.730(0.201)
	MCED2	0.2 & 0.4	0.859	0.730(0.203)

[1] success rate
[2] non-enrichment design
[3] Simon and Simon's enrichment design
[4] Monte Carlo enrichment design using Wald test
[5] Monte Carlo enrichment design using randomization test

Table 4.3: Power and overall success rates from different designs for four equally distributed covariate strata and binary outcomes

Beta	Design	Threshold	Rejection Rate	SR[1](var*N)
(0.5, 0, 1, 0.5, 0, 0, 0, 0)	NED[2]		0.049	0.698(0.219)
	SED[3]		0.034	0.698(1.172)
	MCED1[4]	0.2 & 0.4	0.058	0.698(0.209)
	MCED2[5]	0.2 & 0.4	0.053	0.698(0.219)
(0.5, 0, 1, 0.5, 0, 0.3, 0.3, 0.3)	NED		0.409	0.719(0.206)
	SED		0.380	0.725(0.816)
	MCED1	0.2	0.514	0.720(0.203)
	MCED2	0.2	0.513	0.720(0.200)
	MCED1	0.4	0.653	0.689(0.215)
	MCED2	0.4	0.656	0.689(0.218)
(0.5, 0, 1, 0.5, 0, 0.5, 0.5, 0.5)	NED		0.806	0.731(0.200)
	SED		0.831	0.740(0.509)
	MCED1	0.2	0.892	0.735(0.203)
	MCED2	0.2	0.891	0.735(0.198)
	MCED1	0.4	0.966	0.709(0.216)
	MCED2	0.4	0.964	0.709(0.213)
(0.5, 0, 1, 0.5, 0, 0.7, 0.7, 0.7)	NED		0.970	0.741(0.195)
	SED		0.983	0.753(0.345)
	MCED1	0.2 & 0.4	0.999	0.730(0.215)
	MCED2	0.2 & 0.4	0.999	0.730(0.213)

[1] success rate
[2] non-enrichment design
[3] Simon and Simon's enrichment design
[4] Monte Carlo enrichment design using Wald test
[5] Monte Carlo enrichment design using randomization test

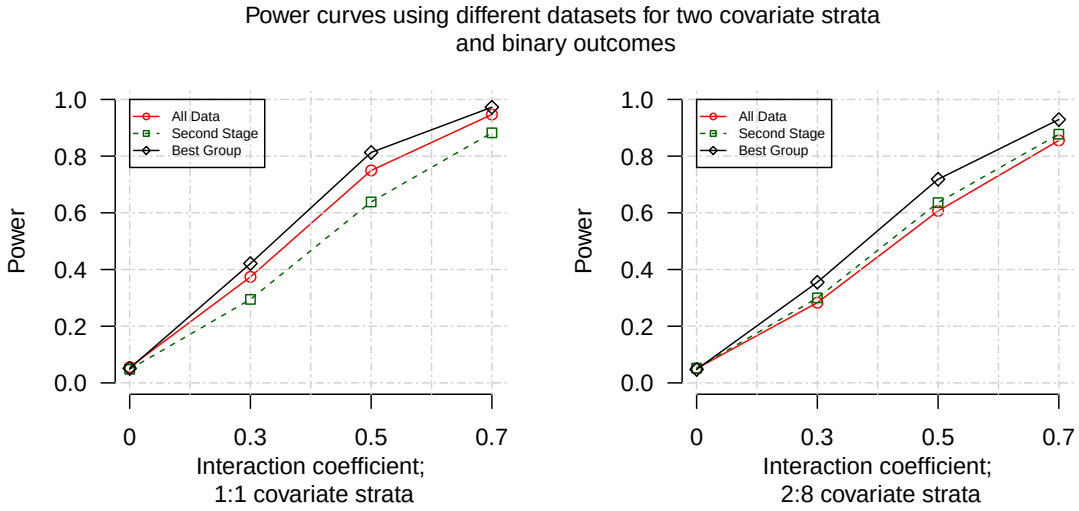

Figure 4.1: Powers using different types of data for binary outcomes

two is used in this section. The test statistic is calculated based on equation 1.14 with $K = 2$. The parameters used in the two covariate strata scenarios are: $\beta_1 = 0.5, \beta_2 = 0, \beta_3 = 0.5$, and β_4 ranging from 0 to 0.7. The sample size of 400 is chosen so that non-enrichment design achieves at least 80% power when $\beta_4 = 0.7$. All four designs preserve type I error rates (Table 4.4 and Table 4.5). Outliers reduce testing powers under all designs. Simon and Simon's enrichment designs are consistently more powerful than non-enrichment designs. Both Monte Carlo enrichment designs are generally more powerful than Simon and Simon's enrichment designs with the exception when there is a relatively small effect size and Monte Carlo enrichment designs using a tighter threshold of 0.2. In this case, since Monte Carlo enrichment designs continue to the second stage with all comers, Simon and Simon's enrichment design is more powerful. A looser threshold of 0.4 leads to an enrichment in the second stage, and thus a higher power.

The parameter values used in the four covariate strata scenarios are: $\beta_1 = 0.5, \beta_2 =$

$0, \beta_3 = 1, \beta_4 = 0.5, \beta_5 = 0$, and $\beta_6 - \beta_8$ ranging from 0 to 0.8. The sample size of 400 is chosen so that the non-enrichment design reaches at least 80% power when $\beta_6 = \beta_7 = \beta_8 = 0.8$ with 10% outliers. All designs preserve type I error rates (Table 4.7 and Table 4.6). The non-enrichment design is consistently less powerful than the other three designs even under Monte Carlo enrichment designs without enrichment in the second stage since Monte Carlo enrichment designs use CARA randomization and assign more patients from the responsive strata to the treatment group. Simon and Simon's enrichment design has higher power when Monte Carlo enrichment designs do not enrich in the second stage under no outliers scenarios. An enriched second stage in a Monte Carlo enrichment design guarantees a higher power. The two Monte Carlo enrichment designs have similar powers under all scenarios. Similar to binary outcomes, using any three types of data preserve type I error rates (Figure 4.2). Using all data yields higher powers than the other two types of data when three out of four strata benefit from the treatment. The power differences are larger when there are outliers and larger effect sizes.

Table 4.4: Type I error rates and powers from different designs for two equally distributed covariate strata and continuous outcomes

Beta	Design	Threshold	Rejection Rate
(0.5, 0, 0.5, 0)	NED[1]		0.048
	SED[2]		0.050
	MCED1[3]	0.2 & 0.4	0.047
	MCED2 [4]	0.2 & 0.4	0.050
(0.5, 0, 0.5, 0.2)	NED		0.160
	SED		0.201
	MCED1	0.2	0.173
	MCED2	0.2	0.172
	MCED1	0.4	0.324
	MCED2	0.4	0.316
(0.5, 0, 0.5, 0.4)	NED		0.462
	SED		0.673
	MCED1	0.2 & 0.4	0.854
	MCED2	0.2 & 0.4	0.853
(0.5, 0, 0.5, 0.6)	NED		0.778
	SED		0.945
	MCED1	0.2 & 0.4	0.994
	MCED2	0.2 & 0.4	0.994

[1] non-enrichment design
[2] Simon and Simon's enrichment design
[3] Monte Carlo enrichment design using Wald test
[4] Monte Carlo enrichment design using randomization test

4.3 Redesigning an existing trial

The National Surgical Adjuvant Breast and Bowel Project (NSABP) B-35 is a phase III trial to compare anastrozole versus tamoxifen in postmenopausal women with hormone (estrogen and/or progesterone) receptor positive ductal carcinoma in situ undergoing lumpectomy plus radiotherapy. The qualified patients were enrolled and randomly assigned $(1 : 1)$ to receive either oral tamoxifen 20 mg per day or oral anastrozole 1 mg per day for 5 years. Margolese et al. (2016) reported the primary results from this study. A total of 3104 patients were enrolled between Jan 1, 2003

Table 4.5: Type I error rates and powers from different designs for two equally distributed covariate strata and continuous outcomes with 10% outliers

Beta	Design	Threshold	Rejection Rate
(0.5, 0, 0.5, 0)	NED[1]		0.047
	SED[2]		0.047
	MCED1[3]	0.2	0.051
	MCED2[4]	0.2	0.048
	MCED1	0.4	0.050
	MCED2	0.4	0.047
(0.5, 0, 0.5, 0.1)	NED		0.053
	SED		0.064
	MCED1	0.2 & 0.4	0.068
	MCED2	0.2 & 0.4	0.068
(0.5, 0, 0.5, 0.3)	NED		0.110
	SED		0.161
	MCED1	0.2 & 0.4	0.222
	MCED2	0.2 & 0.4	0.237
(0.5, 0, 0.5, 0.5)	NED		0.248
	SED		0.373
	MCED1	0.2 & 0.4	0.535
	MCED2	0.2 & 0.4	0.544
(0.5, 0, 0.5, 0.7)	NED		0.431
	SED		0.630
	MCED1	0.2 & 0.4	0.816
	MCED2	0.2 & 0.4	0.826

[1] non-enrichment design

[2] Simon and Simon's enrichment design

[3] Monte Carlo enrichment design using Wald test

[4] Monte Carlo enrichment design using randomization test

Table 4.6: Type I error rates and powers from different designs for four equally distributed covariate strata and continuous outcomes

Beta	Design	Threshold	Rejection Rate
(0.5, 0, 1, 0.5, 0, 0, 0, 0)	NED[1]		0.050
	SED[2]		0.058
	MCED1[3]	0.2 & 0.4	0.040
	MCED2[4]	0.2 & 0.4	0.043
(0.5, 0, 1, 0.5, 0, 0.2, 0.2, 0.2)	NED		0.275
	SED		0.314
	MCED1	0.2 & 0.4	0.285
	MCED2	0.2 & 0.4	0.280
(0.5, 0, 1, 0.5, 0, 0.4, 0.4, 0.4)	NED		0.766
	SED		0.845
	MCED1	0.2	0.827
	MCED2	0.2	0.826
	MCED1	0.4	0.916
	MCED2	0.4	0.923
(0.5, 0, 1, 0.5, 0, 0.6, 0.6, 0.6)	NED		0.978
	SED		0.993
	MCED1	0.2 & 0.4	0.999
	MCED2	0.2 & 0.4	0.999

[1] non-enrichment design
[2] Simon and Simon's enrichment design
[3] Monte Carlo enrichment design using Wald test
[4] Monte Carlo enrichment design using randomization test

Table 4.7: Type I error rates and powers from different designs for four equally distributed covariate strata and continuous outcomes with 10% outliers

Beta	Design	Threshold	Rejection Rate
(0.5, 0, 1, 0.5, 0, 0, 0, 0)	NED[1]		0.048
	SED[2]		0.053
	MCED1[3]	0.2 & 0.4	0.048
	MCED2[4]	0.2 & 0.4	0.042
(0.5, 0, 1, 0.5, 0, 0.2, 0.2, 0.2)	NED		0.108
	SED		0.127
	MCED1	0.2 & 0.4	0.143
	MCED2	0.2 & 0.4	0.152
(0.5, 0, 1, 0.5, 0, 0.4, 0.4, 0.4)	NED		0.334
	SED		0.377
	MCED1	0.2 & 0.4	0.450
	MCED2	0.2 & 0.4	0.472
(0.5, 0, 1, 0.5, 0, 0.6, 0.6, 0.6)	NED		0.627
	SED		0.713
	MCED1	0.2 & 0.4	0.790
	MCED2	0.2 & 0.4	0.808
(0.5, 0, 1, 0.5, 0, 0.8, 0.8, 0.8)	NED		0.856
	SED		0.915
	MCED1	0.2 & 0.4	0.959
	MCED2	0.2 & 0.4	0.965

[1] non-enrichment design

[2] Simon and Simon's enrichment design

[3] Monte Carlo enrichment design using Wald test

[4] Monte Carlo enrichment design using randomization test

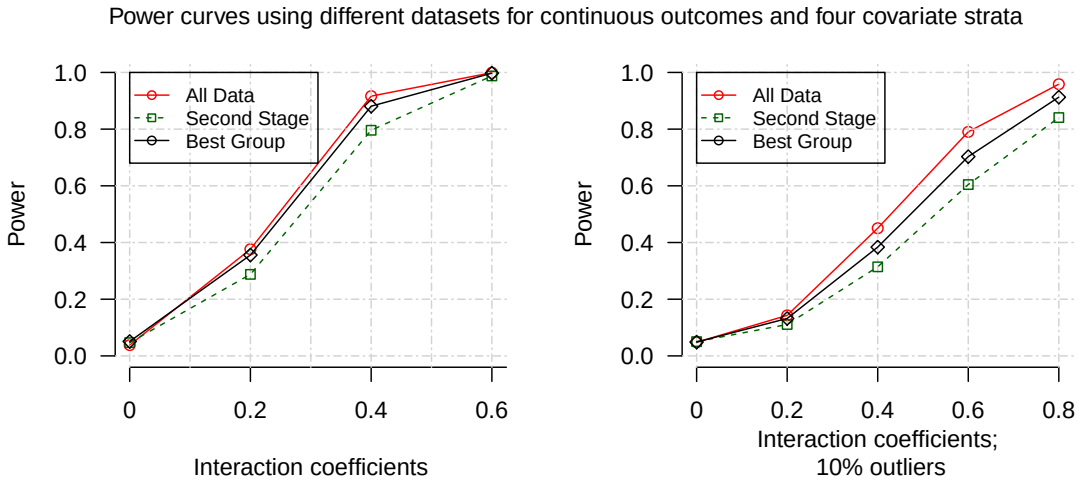

Figure 4.2: Powers using different types of data for continuous outcomes

and Jun 15, 2006 and 3077 patients had disease-free endpoint data by Feb 28, 2015. Anastrozole was found superior to the tamoxifen in the younger than 60-year-old group, but not in the 60 and older group. For those who are younger than 60 years old, 34 out of 724 (4.7%) in the anastrozole group and 63 out of 723 (8.7%) in the tamoxifen group had recurrent breast cancer events. For those who are 60 and older, 56 out of 815 (6.9%) in the anastrozole group and 59 out of 815 (7.2%) in the tamoxifen group had recurrent breast cancer events. Consider a logistic regression model for two treatment groups and one binary covariate with interaction effect:

$$\text{logit}(p_i) = \beta_1 + \beta_2 T_i + \beta_3 Z_i + \beta_4 T_i Z_i, \quad i = 1, \cdots, n,$$

where $T_i = 1$ when a patient is in the anastrozole group, $T_i = 0$ when a patient is in the tamoxifen group, $Z_i = 1$ when a patient is younger than 60-year-old (47%), $Z_i = 0$ when a patient is 60 and older (53%), the outcome $X_i = 1$ when a patient has no recurrent breast cancer events, and $X_i = 0$ when a patient has any recurrent

Table 4.8: Success rates and p-value from different designs based on NSABP trial

Design	p-value	Overall SR[1]	< 60 SR	$>= 60$ SR
NED[2]	0.053	0.933	0.931	0.935
SED[3]	0.082	0.927	0.924	0.932
MCED[4](0.2)	0.208	0.939	0.943	0.935
MCED(0.4)	0.018	0.939	0.941	0.936

[1] success rate
[2] non-enrichment design
[3] Simon and Simon's enrichment design
[4] Monte Carlo enrichment design

breast cancer events. The parameters used in the simulation are: $\beta_1 = 2.5564, \beta_2 = 0.0458, \beta_3 = -0.2055,$ and $\beta_4 = 0.6129$. Under the non-enrichment design, all 3104 patients are equally allocated to the two treatment groups using PBR with block size of 16. Under Simon and Simon's enrichment design, the first 1552 patients are equally allocated using PBR with block size of 16, the enrollment and allocation in the second stage are sequentially performed based on the criterion (equation 1.12). Under the Monte Carlo enrichment design, patients are allocated using DBCD ($\gamma = 2$) with CARA3. A Monte Carlo test with $15,000$ permutations is used after the first 1552 patients are enrolled. Two thresholds (0.2 and 0.4) are tested. A Wald test is used in the final analysis. The Monte Carlo enrichment design with a threshold of 0.4 leads to an enriched second stage with the highest overall success rate and is able to detect a significant treatment effect between two drug groups ($p = 0.0176$). Since both age groups have higher success rates in the anastrozole group, Simon and Simon's enrichment design is enriched after 2070 patients. Monte Carlo enrichment design (0.4) is enriched after 1552 patients. CARA randomization and the enriched second stage lead to 60.3% of patients who are younger than 60 and 54.6% of patients who are 60 and older be allocated to the anastrozole group.

4.4 Conclusion

The proposed two-stage CARA enrichment design using all data from both stages is more powerful and have higher success rate than the traditional non-enrichment design and Simon and Simon's enrichment design when the enrichment is implemented in the second stage. A higher threshold at the interim analysis increases the chance of an enriched second stage. Using a Wald test or a randomization test in the final analysis in the proposed two-stage enrichment design yields similar power levels. Since randomization tests involve much more computation, Wald tests are recommended in the final analysis to test the treatment effect.

Chapter 5: Conclusions and Future Work

In this dissertation, a two-stage enrichment design is proposed. First we use numerical studies to evaluate the performance of different CARA allocation rules for testing the interaction effect. The performance is measured in terms of the testing power and overall success rate. DBCD($\gamma = 2$) targeting CARA2 is the most powerful with the lowest overall success rate among all the CARA procedures compared. DBCD($\gamma = 2$) targeting CARA1 skews the allocation the most in the price of a reduced power. DBCD($\gamma = 2$) targeting CARA3 is chosen since it has a better balance between the power and the overall success rate.

We then propose a Monte Carlo test which uses Monte Carlo resampling of the interaction terms based on regression models to examine the interaction effect. Although the observed and generated score test statistics are used in calculating the p-values, the procedure itself is nonparametric since it is based on the randomization distribution induced by the particular sequence. Population model-based and the proposed tests perform equally well when there are two strata and no misspecified data. However, when there are multiple strata or outliers, the proposed test performs better. It preserves the type I error rate under all scenarios for both binary and continuous outcomes, while inflated type I error rates are observed in the population model-based tests under some situations. It has higher power than the population model-based tests when outliers are presented.

We then use the Monte Carlo test in the interim analysis and develop a two-stage enrichment design. The proposed design is compared with the traditional non-enrichment design and Simon and Simon's design. The proposed design using all

data from both stages is more powerful than the traditional non-enrichment design and Simon's design when an enrichment is implemented in the second stage. Simon's design is more powerful than the proposed design in the scenarios wherein an enrichment is not implemented. A looser threshold increases the chance of an enriched second stage. Simon's design is less powerful than the traditional non-enrichment design when the majority of the patients benefit from the treatment and the effect size is relatively small. Updating the randomization rule with each patient in Simon's binary outcome scenarios potentially leads to reduced sample size and power. Using a randomization test or a Wald test in the final analysis in our design yields similar power levels. A Wald test involves much less computation and is recommended. Although our simulations did not find inflated type I error rates while using the data from the best subgroup only, as discussed in Diao et al. (2018), selecting a best subgroup in the second stage induces biased sampling and leads to an inflated type I error rate if only the best subgroup data are used in the final analysis. Moreover, when multiple subgroups benefit from the treatment, using all data in both stages increases the power.

The proposed design allocates each patient based on patient's covariate profile and all previous patients' responses. In real clinical trials, this could be very costly and practically impossible. A group sequential approach could be considered in the future. When no significant interaction effect is found in the interim analysis, an overall treatment effect could be tested and a futility boundary could be considered to stop the trial earlier. Different numbers of covariate strata can be incorporated into the C code and similar analyses can be performed. However, multiple treatment arms, continuous covariates, survival, and longitudinal outcomes are not addressed. Although in practice, continuous covariates are often being re-categorized into categorical variables.

We reiterate the contribution of this thesis here:

- Compare different CARA procedures and select the CARA procedures that balance efficiency and ethics better for binary and continuous outcomes.
- Propose a Monte Carlo test for testing the treatment-by-covariate interaction effect which can preserve the type I error rate and maintain power under model misspecification.
- Propose a two-stage CARA enrichment design that can preserve the type I error rate, have higher power, and allocate more responsive patients to the better treatment group.

Bibliography

Aletti, G., Ghiglietti, A., Rosenberger, W. F., et al. (2018). Nonparametric covariate-adjusted response-adaptive design based on a functional urn model. *The Annals of Statistics*, 46(6B):3838–3866.

Antognini, A. B. and Zagoraiou, M. (2012). Multi-objective optimal designs in comparative clinical trials with covariates: the reinforced doubly adaptive biased coin design. *The Annals of Statistics*, 40(3):1315–1345.

Ayanlowo, A. and Redden, D. (2008). A two stage conditional power adaptive design adjusting for treatment by covariate interaction. *Contemporary clinical trials*, 29(3):428–438.

Bandyopadhyay, U. and Bhattacharya, R. (2012). An urn based covariate adjusted response adaptive allocation design. *Statistical Methods in Medical Research*, 21(2):135–148.

Bandyopadhyay, U. and Biswas, A. (2001). Adaptive designs for normal responses with prognostic factors. *Biometrika*, 88(2):409–419.

Bandyopadhyay, U., Biswas, A., and Bhattacharya, R. (2007). A covariate adjusted two-stage allocation design for binary responses in randomized clinical trials. *Statistics in medicine*, 26(24):4386–4399.

Biswas, A. and Angers, J.-F. (2002). A bayesian adaptive design in clinical trials for continuous responses. *Statistica neerlandica*, 56(4):400–414.

Biswas, A. and Bhattacharya, R. (2016). A covariate-adjusted response-adaptive allocation for a general class of continuous responses. *Journal of Statistical Theory and Practice*, 10(4):852–863.

Biswas, A., Park, E., and Bhattacharya, R. (2012). Covariate-adjusted response-adaptive designs for longitudinal treatment responses: PEMF trial revisited. *Statistical Methods in Medical Research*, 21(4):379–392.

Chambaz, A., van der Laan, M. J., and Zheng, W. (2014). Targeted covariate-adjusted response-adaptive lasso-based randomized controlled trials. *Modern Adaptive Randomized Clinical Trials: Statistical, Operational, and Regulatory Aspects*, 345–368.

Chang, Y.-c. I. and Park, E. (2013). Sequential estimation for covariate-adjusted response-adaptive designs. *Journal of the Korean Statistical Society*, 42(1):105–116.

Cheung, S. H., Zhang, L.-X., Hu, F., and Chan, W. S. (2014). Covariate-adjusted response-adaptive designs for generalized linear models. *Journal of Statistical Planning and Inference*, 149:152–161.

Cheung, Y. K., Inoue, L. Y., Wathen, J. K., and Thall, P. F. (2006). Continuous bayesian adaptive randomization based on event times with covariates. *Statistics in medicine*, 25(1):55–70.

Diao, G., Dong, J., Zeng, D., Ke, C., Rong, A., and Ibrahim, J. G. (2018). Biomarker threshold adaptive designs for survival endpoints. *Journal of biopharmaceutical statistics*, 28(6):1038–1054.

Dragalin, V. (2006). Adaptive designs: terminology and classification. *Therapeutic Innovation & Regulatory Science*, 40(4):425.

Freidlin, B., McShane, L. M., and Korn, E. L. (2010). Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute*, 102(3):152–160.

Freidlin, B. and Simon, R. (2005). Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research*, 11(21):7872–7878.

Galbete, A. and Rosenberger, W. F. (2016). On the use of randomization tests following adaptive designs. *Journal of biopharmaceutical statistics*, 26(3):466–474.

Hu, F. and Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials*. John Wiley & Sons.

Hu, F., Zhang, L.-X., et al. (2004). Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *The Annals of Statistics*, 32(1):268–301.

Hu, F., Zhang, L.-X., He, X., et al. (2009). Efficient randomized-adaptive designs. *The Annals of Statistics*, 37(5A):2543–2560.

Hu, J., Zhu, H., and Hu, F. (2015). A unified family of covariate-adjusted response-adaptive designs based on efficiency and ethics. *Journal of the American Statistical Association*, 110(509):357–367.

Maitournam, A. and Simon, R. (2005). On the efficiency of targeted clinical trials. *Statistics in Medicine*, 24(3):329–339.

Margolese, R. G., Cecchini, R. S., Julian, T. B., Ganz, P. A., Costantino, J. P., Vallow, L. A., Albain, K. S., Whitworth, P. W., Cianfrocca, M. E., Brufsky, A. M., et al. (2016). Anastrozole versus tamoxifen in postmenopausal women with ductal carcinoma in situ undergoing lumpectomy plus radiotherapy (nsabp b-35): a randomised, double-blind, phase 3 clinical trial. *The Lancet*, 387(10021):849–856.

Mehta, C. R. and Gao, P. (2011). Population enrichment designs: case study of a large multinational trial. *Journal of Biopharmaceutical Statistics*, 21(4):831–845.

Ohwada, S. and Morita, S. (2016). Bayesian adaptive patient enrollment restriction to identify a sensitive subpopulation using a continuous biomarker in a randomized phase 2 trial. *Pharmaceutical Statistics*, 15(5):420–429.

Parhat, P., Rosenberger, W. F., and Diao, G. (2014). Conditional monte carlo randomization tests for regression models. *Statistics in Medicine*, 33(18):3078–3088.

Plamadeala, V., Rosenberger, W. F., et al. (2012). Sequential monitoring with conditional randomization tests. *The Annals of Statistics*, 40(1):30–44.

Renfro, L. A., Coughlin, C. M., Grothey, A. M., and Sargent, D. J. (2014). Adaptive randomized phase ii design for biomarker threshold selection and independent evaluation. *Chinese Clinical Oncology*, 3(1):3.

Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N., and Ricks, M. L. (2001a). Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–913.

Rosenberger, W. F. and Sverdlov, O. (2008). Handling covariates in the design of clinical trials. *Statistical Science*, 23(3):404–419.

Rosenberger, W. F., Sverdlov, O., and Hu, F. (2012). Adaptive randomization for clinical trials. *Journal of Biopharmaceutical Statistics*, 22(4):719–736.

Rosenberger, W. F., Uschner, D., and Wang, Y. (2019). Randomization: The forgotten component of the randomized clinical trial. *Statistics in Medicine*, 38(1):1–12.

Rosenberger, W. F., Vidyashankar, A., and Agarwal, D. K. (2001b). Covariate-adjusted response-adaptive designs for binary response. *Journal of Biopharmaceutical Statistics*, 11(4):227–236.

Russek-Cohen, E. and Simon, R. M. (1997). Evaluating treatments when a gender by treatment interaction may exist. *Statistics in Medicine*, 16(4):455–464.

Simon, N. (2015). Adaptive enrichment designs: applications and challenges. *Clinical Investigation*, 5(4):383–391.

Simon, N. and Simon, R. (2013). Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4):613–625.

Simon, R. and Maitournam, A. (2004). Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*, 10(20):6759–6763.

Spencer, A. V., Harbron, C., Mander, A., Wason, J., and Peers, I. (2016). An adaptive design for updating the threshold value of a continuous biomarker. *Statistics in Medicine*, 35(27):4909–4923.

Still, A. and White, A. (1981). The approximate randomization test as an alternative to the f test in analysis of variance. *British Journal of Mathematical and Statistical Psychology*, 34(2):243–252.

Sverdlov, O., Rosenberger, W. F., and Ryeznik, Y. (2013). Utility of covariate-adjusted response-adaptive randomization in survival trials. *Statistics in Biopharmaceutical Research*, 5(1):38–53.

Wang, S.-J., James Hung, H., and O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal*, 51(2):358–374.

Wang, S.-J., O'Neill, R. T., and Hung, H. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics*, 6(3):227–244.

Yang, B., Zhou, Y., Zhang, L., and Cui, L. (2015). Enrichment design with patient population augmentation. *Contemporary Clinical Trials*, 42:60–67.

Yuan, Y., Huang, X., and Liu, S. (2011). A bayesian response-adaptive covariate-balanced randomization design with application to a leukemia clinical trial. *Statistics in Medicine*, 30(11):1218–1229.

Zhang, L.-X., Hu, F., Cheung, S. H., and Chan, W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics*, 35(3):1166–1182.

Zhang, L.-X. and Hu, F.-f. (2009). A new family of covariate-adjusted response adaptive designs and their properties. *Applied Mathematics-A Journal of Chinese Universities*, 24(1):1–13.

Zhu, H. (2015). Covariate-adjusted response adaptive designs incorporating covariates with and without treatment interactions. *Canadian Journal of Statistics*, 43(4):534–553.

Zhu, H., Hu, F., and Zhao, H. (2013). Adaptive clinical trial designs to detect interaction between treatment and a dichotomous biomarker. *Canadian Journal of Statistics*, 41(3):525–539.

Curriculum Vitae

Li Yang graduated from Central South Univeristy in 2000 with a Bachelor of Medicine degree in clinical medicine. She received her M.S. degree in epidemiology and biostatistics from George Mason University in 2008. She worked as a statistician at Georgetwon University radiology department from 2008 to 2011. She has been working as a statistician at NIH Clinical Center since 2011.