

2020

Novel statistical methods for multi-stage designs in clinical trials with high placebo response

<https://hdl.handle.net/2144/39909>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**NOVEL STATISTICAL METHODS FOR MULTI-STAGE DESIGNS IN
CLINICAL TRIALS WITH HIGH PLACEBO RESPONSE**

by

YUYIN LIU

B.S., Fudan University, 2010
M.S., University of Illinois at Urbana-Champaign, 2012

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2020

© 2020 by
YUYIN LIU
All rights reserved except for Chapter 2,
which is © 2019 Statistics in Medicine
journal, John Wiley & Sons, Inc.

Approved by

First Reader

Gheorghe Doros, Ph.D.
Professor of Biostatistics

Second Reader

Denis Rybin, Ph.D.
Associate Director of Statistics, Pfizer

Third Reader

Timothy C. Heeren, Ph.D.
Professor of Biostatistics

*Dedicated to my beloved husband Shisheng.
This work could not have been completed without his constant encouragement
and inspiration.*

Acknowledgments

At this very last stage of my study at Boston University, I would like to show my appreciation to many people for their continuous support and generous help during this long journey.

First of all, I couldn't adequately express my gratitude to my advisor and friend Prof. Gheorghe Doros, without whom this work would not have been accomplished. From him, I learned the dedication to science and the impeccable work ethic, which always encouraged me to overcome the numerous difficulties, not only from research but also from work and life, over these years.

I also want to say thank you to my great committee members, Drs. Denis Rybin, Timothy Heeren, Michael LaValley, and Howard Cabral, for their thoughtful suggestions throughout the drafting process of my dissertation. I am honored to be a co-author of a publication with them. Some of the joint work is included in this dissertation.

I would like to express the deepest thanks to my colleagues at Baim Institute for Clinical Research. My supervisor Xiaohua Chen always encouraged me to pursue challenges. Without his support, I couldn't imagine I can be a Ph.D. candidate. My dear team members and friends, especially Lanyu, Qi, Yang, and Gerry assisted me a lot in my work and studies, which made it possible for me to accomplish my part-time study at Boston University.

My special thanks to my parents Prof. Liu and Prof. Yu, for their remote support and unconditional love. They are my role models and the starting point of my dream.

NOVEL STATISTICAL METHODS FOR MULTI-STAGE DESIGNS IN CLINICAL TRIALS WITH HIGH PLACEBO RESPONSE

YUYIN LIU

Boston University, Graduate School of Arts and Sciences, 2020

Major Professor: Gheorghe Doros, Ph.D., Professor of Biostatistics

ABSTRACT

Placebo response occurs when a patient perceives an improvement from the psychological effect of receiving treatment rather than from the therapy itself. High placebo response reduces drug-placebo differences and makes it challenging to demonstrate a statistically significant benefit of an active drug over placebo. Two-way enriched design (TED) and sequential enriched design (SED) are two designs to estimate treatment effect with the existence of placebo response. They are extensions of sequential parallel comparison design (SPCD) and have multiple stages with enrichment strategies. This work aims to propose novel analysis methods and evaluate their performances in the framework of TED and SED. TED is a two-stage, randomized, placebo-controlled design with enrichment in ‘placebo non-responders’ and ‘drug responders’. We first consider the placebo non-response as a measurable binary characteristic, either present or absent in an individual. We then discuss the placebo non-response as a characteristic that exists in every subject to a certain degree. We propose to include it in the model as a weight. In addition, we consider placebo non-response and drug non-response as latent characteristics and introduce stochastic components in the classification of the subjects in the setting of TED. SED, as the only three-stage design, aims to exclude subjects who are ‘placebo responders’ and those who never respond to either treatment. Considering the complexity of this design, we critically appraise the performance of SED from different perspectives. We first test the robustness of SED by varying values of parameters. We then calculate the actual sample size and the

proportion of the target population in the sample. We also apply the first two analysis methods to SED. We evaluate these novel methods on a wide range of simulated data scenarios in terms of type I error, mean squared error, and power. From the appraisals above, SED does not benefit from the additional stage. Therefore, in terms of design, we suggest implementing TED rather than SED when placebo response is a critical issue. In terms of the proposed analysis methods, the approach with stochastic components performs the best based on our evaluations, especially when the definition of response is uncertain.

Contents

1	Introduction	1
1.1	Placebo Response	2
1.2	Placebo Lead-in Design	4
1.3	Sequential Parallel Comparison Design	5
1.4	Two-way Enriched Design	7
1.5	Sequential Enriched Design	9
2	Placebo Response as a Binary or Continuous Characteristic in Two-way Enriched Design	11
2.1	Background and Motivation	12
2.2	Methodology	15
2.2.1	Repeated measures model for binary outcomes	15
2.2.2	Repeated measures model for continuous outcomes	17
2.2.3	Weighted repeated measures model for continuous outcomes	22
2.3	Simulation Study	27
2.3.1	Binary outcomes	27
2.3.2	Continuous outcomes	29
2.4	Conclusions	37
3	Treatment Response as Latent Binary Characteristics in Two-way Enriched Design	39
3.1	Background and Motivation	40
3.2	Methodology	43
3.2.1	Response as a latent characteristic	43

3.2.2	Parameter estimation with EM algorithm	46
3.2.3	The definition of treatment effect	48
3.3	Simulation Study	49
3.3.1	Parameter setting	49
3.3.2	Overall performance	53
3.3.3	Comparison with other methods	54
3.3.4	Antidepressant therapy trial example	62
3.4	Conclusions	64
4	Assessment of the Performance of Sequential Enriched Design	67
4.1	Background and Motivation	68
4.2	Methodology	70
4.2.1	Description of population and parameter setting	70
4.2.2	Concerns about the implementation of SED	72
4.2.3	Data generation	75
4.3	Simulation Study	77
4.3.1	Parameter selection	77
4.3.2	Sample size determination	78
4.3.3	Proportion of the target population in the sample	80
4.3.4	New analysis methods	80
4.4	Conclusions	86
5	Summary and Future Studies	91
5.1	Summary	92
5.2	Future Studies	94
	Appendix A M-step in EM Algorithm	95
	Appendix B Formulas for Variance Components in EM Algorithm	100
	Appendix C Robustness Analyses in EM Algorithm	103

Appendix D Parameter Selection in Sequential Enriched Design	126
Appendix E Sample Size Calculation in Sequential Enriched Design	131
Bibliography	134
Curriculum Vitae	138

List of Tables

2.1	Parameter setting for binary outcomes	14
2.2	Type I error for testing $H_0 : \Delta_1 = 0 \cap \Delta_2 = 0 \cap \Delta_3 = 0$	28
2.3	Power for testing $H_0 : \Delta_1 = 0 \cap \Delta_2 = 0 \cap \Delta_3 = 0$	28
3.1	TED parametrization: Distributions	45
3.2	TED parametrization: Mean functions	45
3.3	Simulation parameters	53
3.4	Estimates of the true response probabilities (0.3, 0.3) with $\rho = 0.1$	55
3.5	Departure in Stage I outcome changes in the active drug group	60
3.6	Treatment effect estimates for ADAPT-A trial	64
4.1	Distribution of the overall population	70
4.2	Means of the outcome changes from the baseline to the end of stage	72
4.3	Chen’s parameter selection	73
4.4	Parameters for the mean responses	76
4.5	Proportion estimates of the target population in the sample	81
D.1	Comparison of different designs under Chen’s setting	127
D.2	Comparison of different designs under the alternative setting 1	128
D.3	Comparison of different designs under the alternative setting 2	129
D.4	Comparison of different designs under the alternative setting 3	130
E.1	Sample size at enrollment under Chen’s setting	132
E.2	Sample size at enrollment under the alternative setting 1	132
E.3	Sample size at enrollment under the alternative setting 2	133

List of Figures

1-1	Placebo lead-in design	4
1-2	Sequential parallel comparison design	7
1-3	Two-way enriched design	8
1-4	Sequential enriched design	9
2-1	Repeated measures model parametrization	19
2-2	Weighted repeated measures model parametrization	23
2-3	Type I error within the framework of TED	34
2-4	MSE within the framework of TED	35
2-5	Power within the framework of TED	36
3-1	Two-way enriched design	44
3-2	Type I error assuming correct response threshold and equal treatment effects	56
3-3	MSE assuming correct response threshold and equal treatment effects	58
3-4	Power assuming correct response threshold and equal treatment effects	59
4-1	Sequential enriched design	69
4-2	Type I error within the framework of SED	86
4-3	MSE within the framework of SED	87
4-4	Power within the framework of SED	88
C-1	MSE assuming correct response threshold specification - a fixed Stage II effect in ‘non-responders’ and various Stage II effects in ‘responders’	104
C-2	MSE assuming correct response threshold specification - Stage II effect in ‘responders’ set to half of the Stage II effect in ‘non-responders’	105

C-3	Power assuming correct response threshold specification - Stage II effect in ‘responders’ set to half of the Stage II effect in ‘non-responders’	106
C-4	MSE assuming correct response threshold specification - Stage II effect in ‘responders’ set to two times of the Stage II effect in ‘non-responders’	107
C-5	Power assuming correct response threshold specification - Stage II effect in ‘responders’ set to two times of the Stage II effect in ‘non-responders’	108
C-6	MSE assuming correct response threshold specification - Stage II effect in ‘responders’ set to 0	109
C-7	Power assuming correct response threshold specification - Stage II effect in ‘responders’ set to 0	110
C-8	MSE assuming equal treatment effects but response threshold misspecification	111
C-9	Power assuming equal treatment effects but response threshold misspecification	112
C-10	MSE assuming response threshold misspecification - Stage II effect in ‘responders’ set to half of the Stage II effect in ‘non-responders’	113
C-11	Power assuming response threshold misspecification - Stage II effect in ‘responders’ set to half of the Stage II effect in ‘non-responders’	114
C-12	MSE assuming no ‘placebo-responder’	115
C-13	Power assuming no ‘placebo-responder’	116
C-14	Type I error with $SD_2=6$	117
C-15	Type I error with $SD_2=7$	118
C-16	Type I error with outcome changes in Stage II equal to 80% of that in Stage I	119
C-17	Power with outcome changes in Stage II equal to 80% of that in Stage I	120
C-18	Power with $\delta_3=9.6$ and $\delta_4=19.3$	121
C-19	Power with $\delta_3=8$ and $\delta_4=23.1$	122
C-20	ADAPT-A trial treatment effect estimates	123
C-21	ADAPT-A trial placebo response classification	124
C-22	ADAPT-A trial drug response classification	125

List of Abbreviations

ADT	Antidepressant Therapy
ANCOVA	Analysis of Covariance
CNS	Central Nervous System
EM	Expectation-Maximization
FDA	Food and Drug Administration
GEE	Generalized Estimating Equation
MADRS	Montgomery-Åsberg Depression Rating Scale
MSE	Mean Square Error
OLS	Ordinary Least Square
PhRMA	Pharmaceutical Research and Manufacturers of America
R&D	Research and Development
SED	Sequential Enriched Design
SPCD	Sequential Parallel Comparison Design
TED	Two-way Enriched Design
TSD	Total Standard Deviation

Chapter 1

Introduction

1.1 Placebo Response

In clinical trials, the randomized, double-blinded, placebo-controlled design has been considered a gold standard in the assessment of treatment effect. However, it is well recognized that placebo response can be a critical issue and a reason for failed trials. Placebo response occurs when a sick person perceives an improvement or experiences an improvement in overall health from the psychological effect of receiving treatment rather than from the treatment itself. Many factors influence the generation of placebo response. It may be due to a person's profound desire to get better, increased medical attention as a result of being in an experimental study of a new treatment, or even an unconscious wish by the person to please the physician by getting better.

Placebo response presents in different clinical trials. Some medical fields, such as psychiatry and neuroscience, may be more susceptible to it than others. Over the last few decades, our understanding of the pathophysiology underlying the disease state and dysfunction of the central nervous system (CNS) has tremendously increased; however, treatments for many disorders only provide symptomatic relief [16]. Approximately 9% of new compounds developed for psychiatric disorders are successfully launched to product from Phase I, with a 50% failure rate in efficacy trials in Phase II studies [12, 25].

One challenge for antipsychotic drug development is the pronounced and variable placebo effect observed during clinical trials [31, 32]. According to a landmark systematic review of antidepressant trials in adults conducted by Walsh *et al.* in 2002 [36], the placebo response rate averages 31%, compared to a mean drug response rate of 50%, with a 7% increase per decade from 1981 to 2000, irrespective of the antidepressant comparator. The mean placebo response in recently published clinical schizophrenia trials was 25%, within a range of 0-41% [16].

Due to variable placebo responses (20-70%), nearly 50% of the recent CNS trials failed to show statistical superiority of the active drug over placebo [9, 31]. According to a survey conducted by the Food and Drug Administration (FDA), many placebo-controlled trials in

depression and schizophrenia over a 12-year period (1987–1999) were unable to distinguish between the investigational drug and placebo [17]. The diminished drug-placebo differences and the increased number of failed antipsychotic trials have increased the cost of drug development, delayed the availability of new antipsychotic medications in the market, and also reduced the research spending of pharmaceutical companies in psychotic disorders [1]. According to Pharmaceutical Research and Manufacturers of America (PhRMA), in 2011, only 240 drugs for psychiatric disorders were being developed. This compares to more than 3,000 drugs for cancer and 750 drugs for infectious diseases during the same year [3]. Investments in psychopharmacological drugs have declined by 70% in the last 10 years [24]. Leading pharmaceutical companies withdrew from research and development (R&D) in this area in 2010, largely due to the high risk and cost involved in continuing to invest in this space. The trend was driven largely by cost-cutting initiatives and the desire to scale back and focus on specific aspects of neuroscience [26].

Considering the vicious impact of high placebo response to the trial success rate, physicians, statisticians, and clinical researchers are working to identify the predictors of placebo response and to develop improvements in the conduct of clinical trials to reduce its impact. Doros *et al.* [7] provided a comprehensive summary of factors contributing to placebo response. It may come from patients, investigators/sites, or the study design. Different methods were proposed to control the impact of placebo response, based on the source of it. Through well-conducted randomization and placebo lead-in phase, investigators are trying to reduce the impact of placebo response from patients' expectations. By using dual assessment, centralized ratings, rater drift monitoring, placebo caused by rater bias or measurement error might be well controlled. In addition, several complex designs and analysis methods have been proposed and implemented in the past 20 years, in hopes of increasing the chance of obtaining successful psychiatric clinical trials.

One group of analysis methods is sequential design with enrichment strategies. Enrichment refers to the screening and selecting of patients based on patient characteristics in an

attempt to find a study population in which the effect of a treatment can be most readily demonstrated [5]. Two or three stages are implemented in these designs to identify subjects who would more likely respond to the active drug and less likely respond to placebo. In this way, it is hoped that the actual treatment effect is separated from the placebo response.

1.2 Placebo Lead-in Design

The earliest attempt is to implement a placebo lead-in phase prior to the traditional parallel design, as indicated in Figure 1.1. In placebo lead-in phase, all enrolled subjects are treated with placebo. Usually, an absolute value and/or percent change in a standard scale is pre-specified and used at the end of the placebo lead-in phase to screen out the subjects who are more likely to be ‘placebo responders’. At the beginning of Stage I, only ‘placebo non-responders’ are enrolled in the actual trial and undergo the randomization.

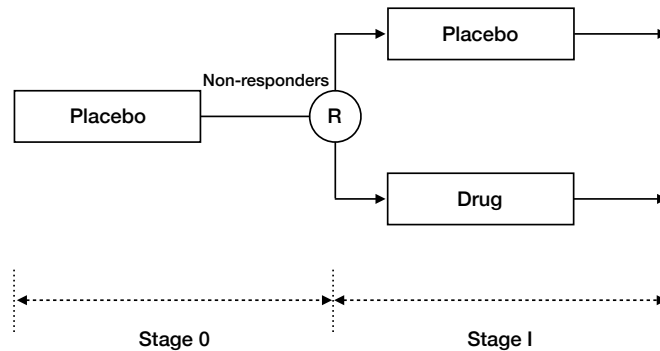


Figure 1.1: Placebo lead-in design

It is expected that the measure of treatment efficacy will be more readily reflected in the subjects who do not respond to placebo. However, through meta-analysis of several published trials, it has been found that this design is not as effective as initially thought,

although it has been widely applied to psychiatric clinical trials [18, 35, 36]. In some earlier meta-analyses [18, 35], antidepressant drug trials with a placebo lead-in phase were compared with those without a placebo lead-in phase. They reached the conclusion that a placebo lead-in phase does not reduce the placebo response rate, increase the drug-placebo difference, or affect the drug response rate. Trivedi and Rush [35] also noted that those who respond to the lead-in may not have the same features as those who ultimately respond to placebo post-randomization, which might also be a reason why the placebo lead-in period does not perform as expected. In Baer and Ivanova’s review [2], they stated the failures of standard placebo lead-in trials might come from short durations of placebo lead-in period or from investigator not being blinded during this period. Attempts have been made, such as the extension of the lead-in phase or the use of double-blinding during the lead-in period, to overcome these deficiencies. The use of an extended lead-in phase might help improving the analysis results, as reported by Chen *et al.* [5] in the meta-analysis led by FDA in 2011. It was noticed that a longer lead-in phase was related to a smaller placebo response rate. However, the use of double-blind placebo lead-in doesn’t necessarily provide a significant treatment difference, even though there is a slightly large drug-placebo difference in placebo non-responders [20, 21].

1.3 Sequential Parallel Comparison Design

In order to overcome the drawback of lack of blindness in placebo lead-in design, the idea of sequential parallel comparison design (SPCD) was first proposed in 2003 [9]. It shares some features with the placebo lead-in design, intending to screen out ‘placebo responders’ in the first stage. Figure 1-2 illustrates the general scheme of SPCD. It includes two stages of equal length. Subjects are randomly assigned to the active drug or placebo at the beginning of Stage I, just as in a traditional parallel design. At the end of Stage I, subjects in the placebo group are classified into ‘placebo responders’ and ‘placebo non-responders’ according to their responses based on the pre-specified criteria. At the end of Stage I,

subjects who do not respond to placebo are re-randomized and participate in Stage II. Subjects who responded to placebo and who are assigned to the active drug in Stage I usually continue with their original treatment, to keep the trial blinded, but their data are not included in the final analysis of treatment effect. Data from all subjects in Stage I and from ‘placebo non-responders’ in Stage II will be used for drug efficacy. SPCD attempts to reduce the effect of placebo response, and at the same time, reduce the required sample size. Unlike conventional randomized placebo-controlled clinical trials, SPCD assigns more subjects to placebo at Stage I, to ensure sufficient subjects entering Stage II [8]. Subject allocation to the active drug group and placebo group in Stage I is determined before the trial, and the sample size for the two groups depends on the expected placebo response. Statistical tests for SPCD with binary outcomes include the likelihood ratio test, Wald test, score test, and linear combination test [9, 11, 13, 34]. For continuous outcomes, a seemingly unrelated regression, ordinary least square estimation, a repeated measures model, and a weighted repeated measures model have been developed, to test and estimate the treatment effect [4, 8, 29, 33].

Two approaches of subject allocation have been proposed. The first one was suggested by Fava *et al.* [9], which arranges subjects into three groups: subjects in Group 1 take placebo for two stages; subjects in Group 2 take placebo in Stage I and switch to the active drug in Stage II; and subjects in Group 3 take the active drug for two stages. The second approach was mentioned in Fava *et al.* [9] and emphasized in Chen *et al.* [4] which includes randomization at the beginning of Stage I and re-randomization at the beginning of Stage II. The second approach is easy to implement. It facilitates the statistical analysis on continuous endpoints since the re-randomization ensures independence between test statistics in Stage I and Stage II [37].

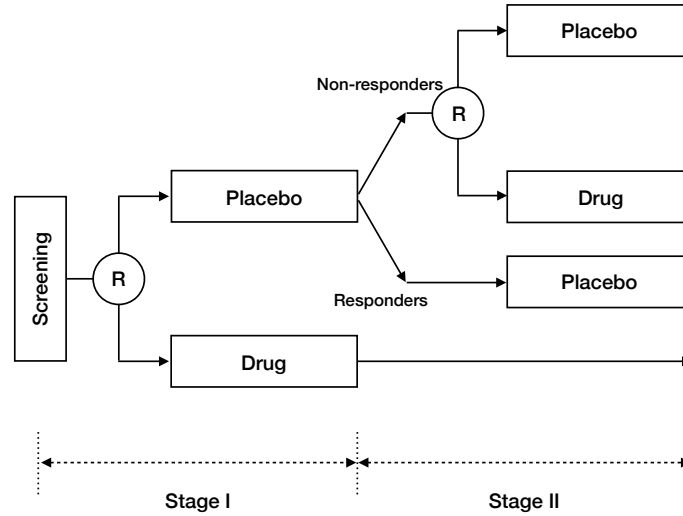


Figure 1·2: Sequential parallel comparison design

1.4 Two-way Enriched Design

The third design called two-way enriched design (TED) [14] was proposed almost 10 years after SPCD was first introduced, as shown in Figure 1·3. TED is an extension of the basic SPCD. Similar to SPCD, TED also consists of two stages. Subjects are randomized to placebo or the active drug at the beginning of Stage I. At the end of Stage I, subjects are classified as ‘placebo responders’, ‘placebo non-responders’, ‘drug responders’, or ‘drug non-responders’ based on pre-determined criteria. ‘Placebo non-responders’ and ‘drug responders’ are re-randomized to placebo or the active drug in Stage II. The subjects who are classified as ‘drug responders’ are examined to see whether the maintenance of response is different between those subjects who switched to placebo versus those remained on the active drug. TED combines the advantages of placebo lead-in and randomized withdrawal. Different from SPCD, in addition to ‘placebo non-responders’, ‘drug responders’ are also included in the analysis of treatment effect, as it is believed that the real treatment effect

should be reflected more obviously in these two subgroups. ‘Placebo responders’ and ‘drug non-responders’ will stay in the study in the second stage with their original treatment assignment, but their data are not included in the final analysis of the treatment effect. Because a portion of the subjects in both the placebo and the active drug groups are to be enrolled into the second stage, a 1:1 randomization scheme is usually applied at the beginning of Stage I, if there is no prior information [14]. Score tests with one, two, and three degrees of freedom were introduced for binary outcomes [14].

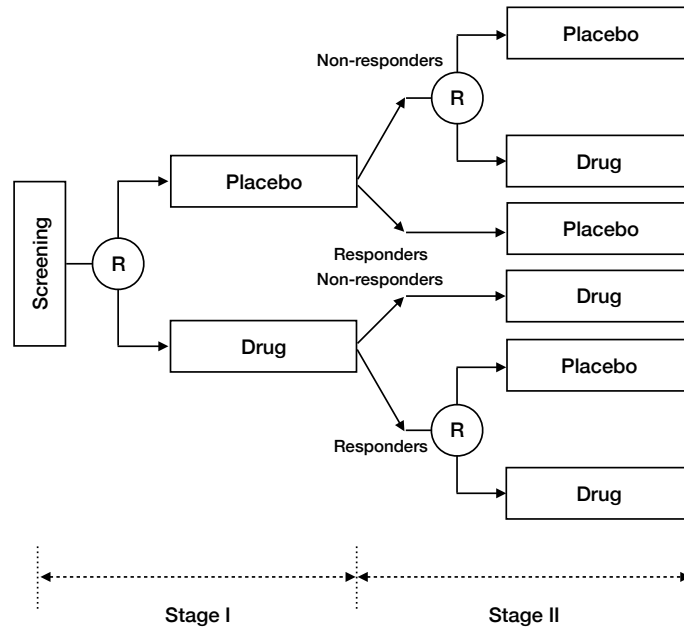


Figure 1.3: Two-way enriched design

There are two approaches to implement TED. The first is to randomize subjects to one of four sequences: placebo-placebo, placebo-drug, drug-placebo, and drug-drug at the beginning of Stage I. The second is to randomize subjects to the active drug or placebo at the beginning of Stage I, and then re-randomize ‘placebo non-responders’ and ‘drug responders’ to the active drug or placebo at the beginning of Stage II. It has been shown

that if the randomization is performed with independent Bernoulli random variables, these two approaches are equivalent [14].

1.5 Sequential Enriched Design

Sequential enriched design (SED) was proposed by Chen *et al.* in 2014 [5], as an extension of SPCD (illustrated in Figure 1.4). However, different from SPCD and TED, this design has three stages. All subjects are treated with placebo in the first stage (Stage 0 - placebo lead-in stage). At the beginning of the second stage (Stage I), ‘placebo non-responders’ are selected and randomized to placebo or the active drug; while ‘placebo responders’ will switch to the active drug for the following stages. At the beginning of the third stage (Stage II), ‘drug responders’ are further selected and re-randomized to placebo or the active drug; while ‘drug non-responders’ will go on with the active drug to the end of the trial.

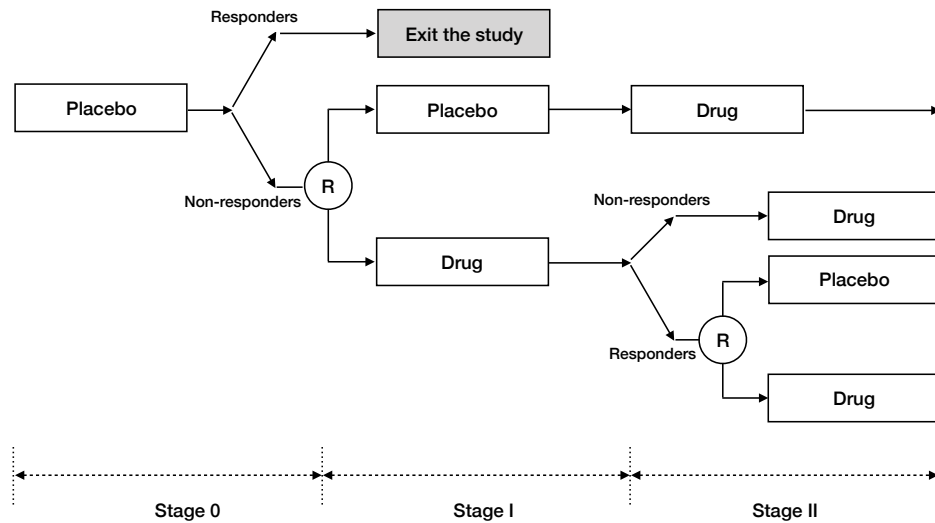


Figure 1.4: Sequential enriched design

The name SED suggests the enrichment is sequential. It first eliminates ‘placebo non-responders’ and then ‘drug non-responders’ within ‘placebo non-responders’. The subjects

who reach the end of the trial should be those who respond to the active drug but do not respond to placebo. Therefore, the stated ultimate goal of SED is to identify the target population, ‘drug responders’ who do not respond to placebo, through the enrichment process by combining the idea of placebo lead-in design, SPCD, and TED. The primary feature of SED is to separate not only patients with high placebo response but also patients who do not respond to the drug from other patients in the study and the analysis, to eliminate the influence of ‘placebo responders’ within the ‘drug responders’ to the purported treatment efficacy [5]. A two-sample t -test was used for analysis and comparison with other designs [5].

Chapter 2

Placebo Response as a Binary or Continuous Characteristic in Two-way Enriched Design

2.1 Background and Motivation

In this chapter, we propose a new approach using a repeated measures model for binary outcomes and two types of repeated measures models for continuous outcomes.

TED has two stages with equal duration. At the end of the first stage, ‘placebo non-responders’ and ‘drug responders’ are selected and re-randomized to enter the second stage. ‘Placebo responders’ and ‘drug non-responders’ remain in the study to maintain blinding. In this design, all the data from the first stage are used for the analysis of the treatment effect. However, only data from ‘placebo non-responders’ and ‘drug responders’ in Stage II are used in this analysis. The rationale for doing so is that it is unlikely to observe a treatment effect in Stage II for ‘placebo responders’ or ‘drug non-responders’. The treatment effect is defined as a weighted average of the outcome difference between the active drug group and the placebo group from three sources: all subjects in the first stage, and ‘placebo non-responders’ and ‘drug responders’ in the second stage. If we assume δ_I as the outcome difference between the two treatment groups in Stage I in the whole population, and δ_{PNR} and δ_{DR} as the outcome difference between the two treatment groups in ‘placebo non-responders’ and ‘drug responders’ in Stage II; then the treatment effect is defined as:

$$\Delta_\omega = \delta_\omega = \omega_1\delta_I + \omega_2\delta_{PNR} + \omega_3\delta_{DR}; \quad \omega_1 + \omega_2 + \omega_3 = 1$$

Ivanova *et al.*[14] proposed a score test with one degree of freedom for the overall treatment effect. They supposed the number of ‘placebo non-responders’ and ‘drug responders’ at the end of Stage I follows a binomial distribution. To facilitate the description of the analysis approach, they used the following notations:

p_1 : P(Drug response in Stage I);

q_1 : P(Placebo response in Stage I);

p_2 : P(Drug response in Stage II | Placebo non-responder in Stage I);

q_2 : P(Placebo response in Stage II | Placebo non-responder in Stage I);

p_3 : P(Drug response in Stage II | Drug responder in Stage I);

q_3 : P(Placebo response in Stage II | Drug responder in Stage I);

s_1 : The probability that a Stage I placebo non-responder continues to the second stage;

s_2 : The probability that a Stage I drug responder continues to the second stage.

Supposing the total sample size to be n with n_1 , n_2 , n_3 , and n_4 subjects assigned to the placebo-placebo group, placebo-drug group, drug-placebo group, and drug-drug group, respectively. Ivanova *et al.* provided Table 2.1 to illustrate the distribution of the subjects in the TED.

The null hypothesis of interest is $H_0 : \delta_\omega = 0$, which indicates no treatment effect either in the first stage or in the second stage. Under the null hypothesis, by setting $\delta_\omega = 0$, Ivanova *et al.* [14] derived the equations for q_1 , q_2 , q_3 , s_1 , and s_2 . They provided a score test statistic with the observed information. The asymptotic distribution of this test statistic under the null hypothesis is a Chi-squared distribution with one degree of freedom. The derivation of the score test statistic can be found in Ivanova *et al.* [14].

It is notable that the original approach is only applicable to binary outcomes, and therefore cannot be used in clinical trials that have continuous outcomes. However, in many clinical trials, the outcome of interest is a continuous variable. This brings up the need for new analysis methods for continuous outcomes. In this chapter, we propose a new approach for binary outcomes and several approaches for continuous outcomes. We perform broad simulations to evaluate the power of these methods and how well they control for type I error.

This chapter consists of three parts. The first part presents the parameterizations and models for the proposed approaches, both for binary and continuous outcomes. The second part presents a large simulation study evaluating the performance of the proposed methods. Finally, in the third part, we discuss the scope of each method and present some general considerations.

Table 2.1: Parameter setting for binary outcomes

Treatment		Response		Count	Probability
Stage I	Stage II	Stage I	Stage II		
Placebo	Placebo (n_1)	No	Yes	n_{11}	$s_1(1 - q_1)q_2$
		No	No	n_{12}	$s_1(1 - q_1)(1 - q_2)$
		Yes	.	n_{13}	q_1
		No	Missing	n_{14}	$(1 - s_1)(1 - q_1)$
Placebo	Drug (n_2)	No	Yes	n_{21}	$s_1(1 - q_1)p_2$
		No	No	n_{22}	$s_1(1 - q_1)(1 - p_2)$
		Yes	.	n_{23}	q_1
		No	Missing	n_{24}	$(1 - s_1)(1 - q_1)$
Drug	Placebo (n_3)	No	.	n_{31}	$1 - p_1$
		Yes	Yes	n_{32}	$s_2p_1q_3$
		Yes	No	n_{33}	$s_2p_1(1 - q_3)$
		Yes	Missing	n_{34}	$(1 - s_2)p_1$
Drug	Drug (n_4)	No	.	n_{41}	$1 - p_1$
		Yes	Yes	n_{42}	$s_2p_1p_3$
		Yes	No	n_{43}	$s_2p_1(1 - p_3)$
		Yes	Missing	n_{44}	$(1 - s_2)p_1$

Note: p_1 : P(Drug response in Stage I); q_1 : P(Placebo response in Stage I); p_2 : P(Drug response in Stage II|Placebo non-responder in Stage I); q_2 : P(Placebo response in Stage II|Placebo non-responder in Stage I); p_3 : P(Drug response in Stage II|Drug responder in Stage I); q_3 : P(Placebo response in Stage II|Drug responder in Stage I). Responses denoted ‘.’ are not included in the analysis by design, n_{14} and n_{24} are ‘placebo non-responders’, and n_{34} and n_{44} are ‘drug responders’, who drop out and do not participate in Stage II [14].

2.2 Methodology

2.2.1 Repeated measures model for binary outcomes

In this design, subjects experience two stages of treatment with equal duration. Their outcomes are measured three times: at baseline, the end of Stage I, and the end of Stage II. It is natural to consider the outcomes as repeated measures for a single subject. We still use the notation from Ivanova's manuscript. Without loss of generality, we assume s_1 and s_2 are both equal to 1. That is, all 'placebo non-responders' and 'drug responders' continue to the second stage. This restriction can be easily removed. Logistic regression is used to model the response probabilities in Stage I and Stage II, in 'placebo non-responders' and 'drug responders', respectively. The model can be specified as follows:

At Stage I:

$$\text{Equation (1): } \text{logit}(p_{1i}) = \alpha_{01} + \delta_1 G_{1i}; \quad i = 1 : N$$

where $p_{1i} = p_1$ if the subject receives the active drug in Stage I and $p_{1i} = q_1$ if the subject receives placebo in Stage I. G_{1i} is the treatment indicator for the data during Stage I. N is the total number of subjects enrolled into the study.

At Stage II:

$$\text{Equation (2): } \text{logit}(p_{2i}) = \alpha_{02} + \delta_2 G_{2i}; \quad i = 1 : n_{PNR}$$

$$\text{Equation (3): } \text{logit}(p_{2i}) = \alpha_{03} + \delta_3 G_{2i}; \quad i = (n_{PNR} + 1) : (n_{PNR} + n_{DR})$$

$$\text{Equation (4): } \text{logit}(p_{2i}) = \alpha_{04} + \delta_4 G_{2i}; \quad i = (n_{PNR} + n_{DR} + 1) : N$$

The first two equations relate the outcome at the end of Stage II to the outcome at the end of Stage I and treatment allocation during the second stage for 'placebo non-responders' and 'drug responders'. G_{2i} is the treatment indicator for the data during Stage II. In Equation (2), $p_{2i} = p_2$ if the 'placebo non-responder' receives the active drug in Stage

II and $p_{2i} = q_2$ if the ‘placebo non-responder’ receives placebo in Stage II. n_{PNR} is the number of ‘placebo non-responders’ entering Stage II. In Equation (3), $p_{2i} = p_3$ if the ‘drug responder’ receives the active drug in Stage II and $p_{2i} = q_3$ if the ‘drug responder’ receives placebo in Stage II. n_{DR} is the number of ‘drug responders’ entering Stage II. In Equation (4), $G_{2i} = 1$ for ‘drug non-responders’ and $G_{2i} = 0$ for ‘placebo responders’.

In matrix form, then the logistic regression models can be written as:

$$\log\left(\frac{p_{ji}}{1 - p_{ji}}\right) = \mathbf{x}'_{ji}\boldsymbol{\delta}; \quad i = 1 : N, j = 1, 2$$

where \mathbf{x}_{ji} is the covariate matrix for i^{th} subject on the j^{th} treatment. When $j = 1$, it is the treatment in Stage I; when $j = 2$, it is the treatment in Stage II. As there is a correlation between the outcomes in the two stages within the same subject, we will use generalized estimating equations (GEE) to estimate the treatment effects. The generalized estimating equation for estimating the regression parameters $\boldsymbol{\delta}$ is an extension of the independence estimating equation to correlated data. Given the covariance structure \mathbf{V}_i , the parameters $\boldsymbol{\delta}$ are estimated by solving

$$S(\boldsymbol{\delta}) = \sum_{i=1}^N \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{p}_i(\boldsymbol{\delta})) = 0$$

where $\mathbf{D}_i' = \frac{\partial \mathbf{p}_i}{\partial \boldsymbol{\delta}}$. Since $g(p_{ji}) = \mathbf{x}'_{ji}\boldsymbol{\delta}$, where g is the link function, the partial derivatives of the mean with respect to the regression parameters for the i^{th} subject is given by

$$\mathbf{D}_i' = \frac{\partial \mathbf{p}_i}{\partial \boldsymbol{\delta}} = \begin{bmatrix} \frac{x_{1i1}}{g'(p_{1i})} & \frac{x_{2i1}}{g'(p_{2i})} \\ \frac{x_{1i2}}{g'(p_{1i})} & \frac{x_{2i2}}{g'(p_{2i})} \end{bmatrix}$$

The variance structure is chosen to improve the efficiency of the parameter estimates. As we don't have prior information about the covariance structure, we will assume an unstructured covariance matrix:

$$\text{Corr}(Y_{ji}, Y_{ki}) = \begin{cases} 1, & \text{if } j = k \\ a_{jk}, & \text{if } j \neq k \end{cases}$$

The Newton-Raphson algorithm is used to obtain the final estimate of the parameters. As mentioned earlier, 'placebo responders' and 'drug non-responders' are not used for the analysis of the treatment effect. Let $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3, 0)$, and then the estimated treatment effect can be specified as:

$$\delta_\omega = \boldsymbol{\omega}' \boldsymbol{\delta} = \omega_1 \delta_1 + \omega_2 \delta_2 + \omega_3 \delta_3; \quad \omega_1 + \omega_2 + \omega_3 = 1$$

which represents the weighted average treatment effect in all subjects in Stage I, and in 'placebo non-responders' and 'drug responders' in Stage II. The hypotheses to be tested can be specified as $H_0 : \delta_\omega = 0$ vs. $H_1 : \delta_\omega \neq 0$. Let $\hat{\boldsymbol{\delta}}$ be the estimated regression parameters resulting from solving the GEE under the restricted model $\boldsymbol{\omega}' \boldsymbol{\delta} = 0$, and let $S(\hat{\boldsymbol{\delta}})$ be the generalized estimating equation at $\hat{\boldsymbol{\delta}}$, then the test statistic can be written as:

$$T = S(\hat{\boldsymbol{\delta}})' \boldsymbol{\Sigma}_m \boldsymbol{\omega} (\boldsymbol{\omega}' \boldsymbol{\Sigma}_e \boldsymbol{\omega})^{-1} \boldsymbol{\omega}' \boldsymbol{\Sigma}_m S(\hat{\boldsymbol{\delta}})$$

where $\boldsymbol{\Sigma}_m$ is the model-based covariance estimate and $\boldsymbol{\Sigma}_e$ is the empirical covariance estimate. Then the statistic T will follow approximately a Wald Chi-squared test with 1 degree of freedom under the null hypothesis.

2.2.2 Repeated measures model for continuous outcomes

In the original analytical approach, the authors suppose the outcome is a binary variable. However, in many clinical trials, the outcome of interest is a continuous variable. The

following methods are for the analysis of continuous outcomes.

Ivanova *et al.* [14] suggested that if no information is available regarding true response rates, it is better to assign equal numbers of subjects to the active drug and placebo in Stage I. We assume that in Stage I the subjects are randomized 1:1 to placebo or the active drug, whereas in Stage II ‘placebo non-responders’ and ‘drug responders’ are randomized 1:1 to placebo or the active drug. ‘Placebo responders’ continue with placebo in Stage II, and ‘drug non-responders’ go on with the active drug in Stage II. Figure 2.1 shows a graphical representation of the effects estimated with a repeated measures model. In Stage I, the horizontal line indicates the level of mean outcome in the whole population at baseline. The distances between the horizontal line and the dashed lines present the changes in outcome from baseline to the end of Stage I. In Stage II, the horizontal lines indicate the levels of mean outcome in ‘placebo non-responders’, ‘placebo responders’, ‘drug non-responders’ and ‘drug responders’, respectively, at the beginning of Stage II. The distances between the horizontal lines and the dashed lines present the changes in outcome from the end of Stage I to the end of Stage II. As we are looking at an adverse outcome, lower scores indicate improvement.

In the figure,

δ_0 : The mean change in outcome from baseline to the end of Stage I in placebo subjects;

δ_1 : The difference in the mean change in outcome from baseline to the end of Stage I between subjects in the active drug group and in the placebo group (treatment effect in Stage I);

δ_{01} : The difference in the mean outcome at the end of Stage I between all placebo subjects and ‘placebo non-responders’;

δ_{02} : The difference in the mean outcome at the end of Stage I between all placebo subjects and ‘placebo responders’;

δ_{11} : The difference in the mean outcome at the end of Stage I between all drug subjects and ‘drug non-responders’;

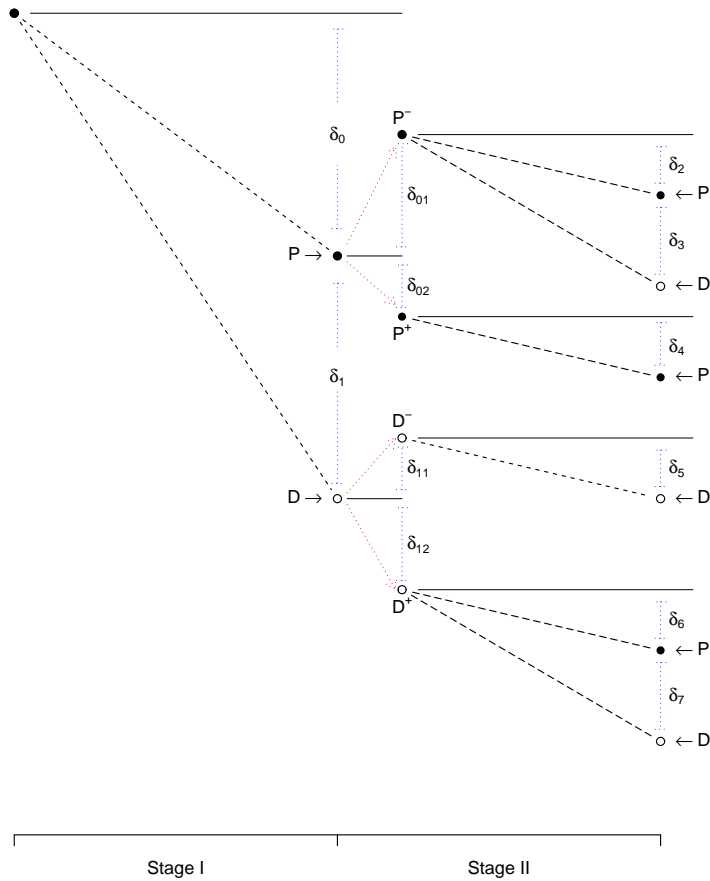


Figure 2.1: Repeated measures model parametrization

δ_{12} : The difference in the mean outcome at the end of Stage I between all drug subjects and ‘drug responders’;

δ_2 : The mean change in outcome from the beginning to the end of Stage II in ‘placebo non-responders’ who were randomized to placebo in Stage II;

δ_3 : The difference in the mean change in outcome from the beginning to the end of Stage II between ‘placebo non-responders’ who are randomized to the active drug and who are randomized to placebo in Stage II (treatment effect part I in Stage II);

δ_4 : The mean change in outcome from the beginning to the end of Stage II in ‘placebo responders’;

δ_5 : The mean change in outcome from the beginning to the end of Stage II in ‘drug non-responders’;

δ_6 : The mean change in outcome from the beginning to the end of Stage II in ‘drug responders’ who were randomized to placebo in Stage II;

δ_7 : The difference in the mean change in outcome from the beginning to the end of Stage II between ‘drug responders’ who are randomized to the active drug and who are randomized to placebo in Stage II (treatment effect part II in Stage II).

In the proposed repeated measures model, we will use all available data collected at baseline, the end of Stage I, and the end of Stage II. However, the treatment effect will be estimated with outcomes at baseline, at the end of Stage I and outcomes from ‘placebo non-responders’ and ‘drug responders’ at the end of Stage II. In the following, we will list all the models that are used in this analysis approach. Linear regression will be used to model the outcomes and the correlation between stages.

At Stage I:

$$\text{Equation (1): } \Delta Y_{1i} = \alpha_{01} + \alpha_{11}Y_{1i} + \delta_1 G_{1i} + \epsilon_{1i}; \quad i = 1 : N$$

where ΔY_{1i} is the change in outcome from baseline to the end of Stage I. G_{1i} is the treatment indicator in Stage I.

At Stage II:

$$\text{Equation (2): } \Delta Y_{2i} = \alpha_{02} + \alpha_{12}Y_{2i} + \delta_3 G_{2i} + \epsilon_{2i}; \quad i = 1 : n_{PNR}$$

$$\text{Equation (3): } \Delta Y_{2i} = \alpha_{03} + \alpha_{13}Y_{2i} + \delta_7 G_{2i} + \epsilon_{3i}; \quad i = (n_{PNR} + 1) : (n_{PNR} + n_{DR})$$

$$\text{Equation (4): } \Delta Y_{2i} = \alpha_{04} + \alpha_{14}Y_{2i} + \epsilon_{4i}; \quad i = (n_{PNR} + n_{DR} + 1) : (n_{PNR} + n_{DR} + n_{PR})$$

$$\text{Equation (5): } \Delta Y_{2i} = \alpha_{05} + \alpha_{15}Y_{2i} + \epsilon_{5i}; \quad i = (n_{PNR} + n_{DR} + n_{PR} + 1) : N$$

The first two equations relate the change in the outcome from the end of Stage I to the end of Stage II (ΔY_{2i}) to the outcome at the end of Stage I (Y_{2i}) and the new treatment

assignment (G_{2i}) for ‘placebo non-responders’ and ‘drug responders’, respectively. The last two equations present the relationship between the change in outcome from the end of Stage I to the end of Stage II and the outcome at the end of Stage I for ‘placebo responders’ and ‘drug non-responders’.

It is assumed that the error terms $\{\epsilon_{1i}\}$, $\{\epsilon_{2i}\}$, $\{\epsilon_{3i}\}$, $\{\epsilon_{4i}\}$, and $\{\epsilon_{5i}\}$ are independently and identically distributed across individuals. As subjects have outcomes recorded for both of the stages, the outcomes in the two stages are correlated within each subject. We assume the correlation is the same across the subjects, regardless of whether their data in the second stage are used for the final estimate of the treatment effect. Therefore, we have the following covariance matrices:

$$(\epsilon_{1i}, \epsilon_{ji}) \sim N \left[(0, 0), \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right]; \quad i = 1 : N, \quad j = 2 : 5$$

Therefore, the contrast of interest is

$$\delta_\omega = \omega_1 \delta_1 + \omega_2 \delta_3 + \omega_3 \delta_7; \quad \omega_1 + \omega_2 + \omega_3 = 1$$

which represents the weighted average treatment effect in all subjects in Stage I and in ‘placebo non-responders’ and ‘drug responders’ in Stage II. This is estimated by $\hat{\delta}_\omega = \omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_3 + \omega_3 \hat{\delta}_7$, with $\hat{\delta}_1$, $\hat{\delta}_3$, and $\hat{\delta}_7$, the model-based estimates of δ_1 , δ_3 , and δ_7 , respectively. A test for $H_0 : \delta_\omega = 0$ is based on the test statistic

$$T = \frac{\omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_3 + \omega_3 \hat{\delta}_7}{\sqrt{\omega_1^2 \text{Var}(\hat{\delta}_1) + \omega_2^2 \text{Var}(\hat{\delta}_3) + \omega_3^2 \text{Var}(\hat{\delta}_7) + 2\omega_1 \omega_2 \text{Cov}(\hat{\delta}_1, \hat{\delta}_3) + 2\omega_1 \omega_3 \text{Cov}(\hat{\delta}_1, \hat{\delta}_7) + 2\omega_2 \omega_3 \text{Cov}(\hat{\delta}_3, \hat{\delta}_7)}}$$

where the variances and covariances can be derived from the model specified ahead. It is assumed that T will follow approximately the standard normal distribution under the null hypothesis H_0 .

2.2.3 Weighted repeated measures model for continuous outcomes

Rybin *et al.* [29] proposed a weighted repeated measures model that uses all the data in the placebo group, instead of only including ‘placebo non-responders’ in the estimation of treatment effect. In this section, we try to apply this idea to TED. Figure 2.3 shows a graphical representation of the effects estimated with the weighted repeated measures model. This figure is basically the same as the one for repeated measures model, with the only difference being that both ‘placebo responders’ and ‘placebo non-responders’ are re-randomized at the end of Stage I in a 1:1 ratio to the active drug or placebo in Stage II. The treatment effect in the placebo group is split into two parts, that in ‘placebo non-responders’ (δ_{31}) and that in ‘placebo responders’ (δ_{32}).

In Doros’ repeated measures model, placebo non-response is defined as a binary status, which is determined through the mean change in outcome from baseline to the end of Stage I. However, it is subjective to classify subjects as ‘placebo non-responders’ using only a single value. It also may lead to a high misclassification rate with only two groups, ‘placebo responders’ and ‘placebo non-responders’. In order to solve this problem, Rybin *et al.* [29] described the placebo non-response as a subject characteristic with a zero-to-one scale. That is, if the subject is more like a ‘placebo non-responder’, the subject will get a value closer to one; if the subject tends to be a ‘placebo responder’, the value will be closer to zero. This scale is used in the repeated measures model as a weight. Therefore, subjects more like ‘placebo non-responders’ will have larger weights in the model to estimate the treatment effect. These weights are specific to assessing placebo response only; we do not apply a similar characteristic to the subjects in the active drug group, because it is impossible to observe placebo response in the subjects initially assigned to the active drug group. The full model can be specified as the following four components:

At Stage I:

$$\text{Equation (1): } \Delta Y_{1i} = \alpha_{01} + \alpha_{11}Y_{1i} + \delta_1 G_{1i} + \epsilon_{1i}; \quad i = 1 : N$$

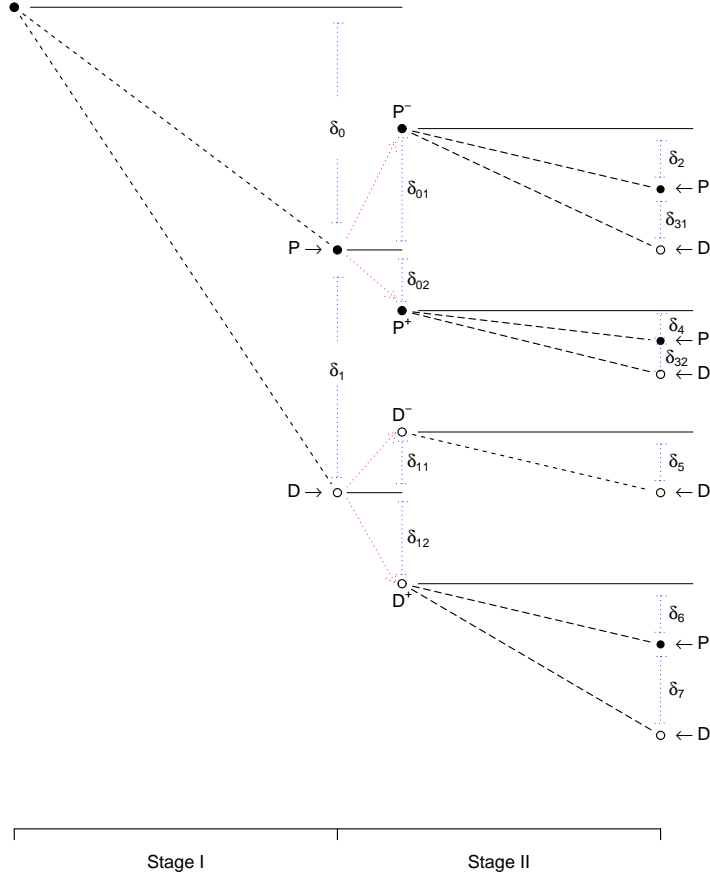


Figure 2-2: Weighted repeated measures model parametrization

At Stage II:

Equation (2): $\Delta Y_{2i} = \alpha_{02} + \alpha_{12}Y_{2i} + \delta_3 G_{2i} + \epsilon_{2i}; \quad i = 1 : n_P$

Equation (3): $\Delta Y_{2i} = \alpha_{03} + \alpha_{13}Y_{2i} + \delta_7 G_{2i} + \epsilon_{3i}; \quad i = (n_P + 1) : (n_P + n_{DR})$

Equation (4): $\Delta Y_{2i} = \alpha_{04} + \alpha_{14}Y_{2i} + \epsilon_{4i}; \quad i = (n_P + n_{DR} + 1) : N$

The first equation relates the change in the outcome from baseline to the end of Stage I (ΔY_{1i}) to the outcome at baseline (Y_{1i}) and the treatment assignment during Stage I (G_{1i}). The second equation and the third equation relate the change in the outcome from the end of Stage I to the end of Stage II (ΔY_{2i}) to the outcome at the end of Stage I (Y_{2i}) and the

new treatment assignment (G_{2i}) for the placebo group and ‘drug responders’ in Stage II. Finally, the last equation relates the change in the outcome from the end of Stage I to the end of Stage II (ΔY_{2i}) to the outcome at the end of Stage I (Y_{2i}) for ‘drug non-responders’ in Stage II.

Next, we specify the covariance matrix for this model. As the outcome is normally distributed, the errors ϵ in the model are also normally distributed with mean $E(\epsilon) = \mathbf{0}$ and variance $\text{Var}(\epsilon) = \sigma^2 \mathbf{\Sigma}$, where σ^2 is unknown, and $\mathbf{\Sigma}$ is defined as follows, reflecting the correlation between Stage I and Stage II (ρ_{12}).

$$\mathbf{\Sigma}_i = v_i^{-1/2} \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix} v_i^{-1/2}; \quad i = 1 : n_P$$

$$\mathbf{\Sigma}_i = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix} \quad i = n_P : N$$

The weights are set to 1 for all subjects in Stage I. In Stage II, all subjects in the active drug group in Stage I are assigned to a weight of 1, but subjects in the placebo group in Stage I are assigned weights based on their non-response to placebo. Therefore, v_i takes the following form:

$$\mathbf{v}_i = \begin{bmatrix} 1 & 0 \\ 0 & v_i \end{bmatrix} \quad i = 1 : n_P$$

In matrix form, the model can be written as $\mathbf{\Delta Y}_i = \mathbf{X}_i \boldsymbol{\delta} + \epsilon_i$, where $\mathbf{\Delta Y}_i$ is a vector of outcomes and \mathbf{X}_i is the covariate matrix for individual i . The generalized least squares estimate for the vector of coefficients is $\hat{\boldsymbol{\delta}} = \{\sum_{i=1}^N \mathbf{X}_i' \mathbf{\Sigma}_i^{-1} \mathbf{X}_i\}^{-1} \sum_{i=1}^N \mathbf{X}_i' \mathbf{\Sigma}_i^{-1} \mathbf{\Delta Y}_i$. With $\mathbf{\Sigma}_i$ known, the variance of the estimate is $\text{Var}(\hat{\boldsymbol{\delta}}) = \sigma^2 \{\sum_{i=1}^N \mathbf{X}_i' \mathbf{\Sigma}_i^{-1} \mathbf{X}_i\}^{-1}$.

Therefore, the contrast of interest can be specified as:

$$\delta_\omega = \omega_1 \delta_1 + \omega_2 \delta_3 + \omega_3 \delta_7; \quad \omega_1 + \omega_2 + \omega_3 = 1$$

It is estimated by $\hat{\delta}_\omega = \omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_3 + \omega_3 \hat{\delta}_7$, with $\hat{\delta}_1$, $\hat{\delta}_3$, and $\hat{\delta}_7$, the model-based estimates

of δ_1 , δ_3 , and δ_7 , respectively. A test for $H_0 : \delta_\omega = 0$ is based on the test statistic

$$T = \frac{\omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_3 + \omega_3 \hat{\delta}_7}{\sqrt{\omega_1^2 \text{Var}(\hat{\delta}_1) + \omega_2^2 \text{Var}(\hat{\delta}_3) + \omega_3^2 \text{Var}(\hat{\delta}_7) + 2\omega_1 \omega_2 \text{Cov}(\hat{\delta}_1, \hat{\delta}_3) + 2\omega_1 \omega_3 \text{Cov}(\hat{\delta}_1, \hat{\delta}_7) + 2\omega_2 \omega_3 \text{Cov}(\hat{\delta}_3, \hat{\delta}_7)}}$$

where the treatment effects, variances, and covariances are estimated from the model specified subsequently. It is assumed that T follows approximately the standard normal distribution under the null hypothesis.

In order to get the zero-to-one scale of placebo non-response, many options are available. Propensity score is one of the options [28]. A logistic regression is constructed with baseline characteristics, and the outcomes at both baseline and at the end of Stage I. The probability of being a ‘placebo non-responder’ is estimated through the model:

$$\text{logit}(p_{R_i=0}) = \alpha + \beta_0 Y_{1i} + \beta_1 X_i; \quad i = 1 : n_P$$

where Y_{1i} is the outcome at baseline for subject i , X_i is another baseline characteristic, R_i is a response indicator with a value of 0 or 1 (for ‘placebo non-responder’ or ‘placebo responder’, respectively), and n_P is the number of subjects in the placebo group at Stage I. However, this approach can only yield a pseudo weight, as binary placebo response status is still needed to construct the model. In the next section, another method is introduced, which provides a more flexible and realistic method of analysis.

K-means clustering is another option. It is obvious that ‘placebo non-responders’ will have some characteristics in common. This similarity in some dimensions allows for the possibility of grouping them together. Rybin *et al.* [29] proposed K-means clustering to determine how to classify subjects as ‘placebo responders’ and ‘placebo non-responders’. Suppose that we collect n baseline characteristics of subjects and that each subject can be located with a point in the R^n space. The relative distance between the subjects helps us to group them together. In our case, we will only need two groups: one is for ‘placebo responders’, and the other is for ‘placebo non-responders’.

In this chapter, two variables are used for clustering: the percent change from baseline to the end of Stage I, and the baseline value in Stage II. It is expected that ‘placebo non-responders’ would have a low percent change from baseline in Stage I by definition of non-response, and would have a high baseline value in Stage II due to the smaller change in outcome in Stage I. These two items are likely correlated. Therefore, two principal components will be constructed to preserve the relative distance based on the Euclidean distance. The following procedures are used to determine the weights for ‘placebo responders’ and ‘placebo non-responders’:

1. K-means clustering (with $K=2$) is performed on the two principal components. The centers of clusters, the variability within clusters, and the total variability are retrieved from the analysis.
2. The center-point coordinates, adjusted for within-cluster variability, are computed as follows:

$$c_1 = \frac{m_{11}s_{21} + m_{21}s_{11}}{s_{11} + s_{21}}$$

$$c_2 = \frac{m_{12}s_{22} + m_{22}s_{12}}{s_{12} + s_{22}}$$

3. The distance d_i to the center-point for subject i is computed as follows:

$$d_i = (-1)^{C_i} \sqrt{(p_{1i} - c_{1i})^2 + (p_{2i} - c_{2i})^2}; \quad i = 1 : n_P$$

where $C_i \in 1, 2$ is the cluster, and p_{1i} and p_{2i} are the two principal components for subject i .

4. The subject-specific scores are then derived as: $w_{k,i} = \Phi_k(d_i)$ where Φ_k is the cumulative distribution function of the normal distribution with mean 0 and standard deviation $k \times \text{TSD}$ (TSD is total standard deviation determined in K-means clustering step).

In this approach, the weights are set to 1 for all subjects in Stage I. In Stage II, all subjects

in the active drug group at Stage I are assigned to a weight of 1, but subjects in the placebo group in Stage I are assigned weights based on the scores derived from the steps above.

2.3 Simulation Study

2.3.1 Binary outcomes

An abroad simulation was undertaken to evaluate the performance of the GEE method. Data were generated from binomial distribution by using different pairs of parameters p_1 , q_1 , p_2 , q_2 , p_3 , and q_3 . Without loss of generality, both s_1 and s_2 were set to 1, so all ‘placebo non-responders’ and ‘drug responders’ entered the second stage. For each scenario, we generated 1,000,000 data sets for the evaluation of type I error and 10,000 data sets for power comparison between the proposed method and the score test.

Type I errors are evaluated under nine different combinations of response probabilities (Table 2.2). In the first six scenarios, the response probability in ‘placebo non-responders’ remains the same, while those in Stage I and ‘drug responders’ change from case to case. In the last three scenarios, the response probabilities in Stage I and ‘drug responders’ remain the same and enable the change in response probability in ‘placebo non-responders’. Type I errors are presented under different ω_1 . As we assume ‘placebo non-responders’ play the same role as ‘drug responders’, we applied the same weights for ‘placebo non-responders’ and ‘drug responders’ in Stage II. That is $\omega_2 = \omega_3$. For most of the scenarios, GEE has a better performance than the score test. The type I error is well controlled under 0.05, except in a few cases.

In the evaluation of power, we chose the same parameters as presented by Ivanova and Tamura. In Table 2.3, nine scenarios are listed. N was provided by Ivanova *et al.* in their manuscript, which is the number of subjects needed to reach the targeted power in the score test. In comparison, we used the same sample size to get the power for the GEE method. In scenario 1 and scenario 5, the effect size in the first stage is the same as that in the second stage. In scenario 2 and scenario 3, the effect size is larger in the first stage and in

Table 2.2: Type I error for testing $H_0 : \Delta_1 = 0 \cap \Delta_2 = 0 \cap \Delta_3 = 0$

p_1	q_1	p_2	q_2	p_3	q_3	Score Test	GEE $\omega_1=0.5$	GEE $\omega_1=0.6$	GEE $\omega_1=0.7$	GEE $\omega_1=0.8$	GEE $\omega_1=0.9$
0.4	0.4	0.4	0.4	0.9	0.9	0.053	0.045	0.047	0.050	0.050	0.049
0.5	0.5	0.4	0.4	0.9	0.9	0.049	0.048	0.047	0.049	0.051	0.047
0.3	0.3	0.4	0.4	0.9	0.9	0.053	0.046	0.048	0.047	0.048	0.047
0.4	0.4	0.4	0.4	0.7	0.7	0.048	0.048	0.052	0.050	0.053	0.053
0.5	0.5	0.4	0.4	0.7	0.7	0.048	0.048	0.049	0.054	0.049	0.049
0.3	0.3	0.4	0.4	0.7	0.7	0.052	0.050	0.052	0.050	0.049	0.051
0.4	0.4	0.4	0.4	0.8	0.8	0.051	0.048	0.051	0.052	0.050	0.048
0.4	0.4	0.5	0.5	0.8	0.8	0.050	0.049	0.049	0.053	0.046	0.050
0.4	0.4	0.3	0.3	0.8	0.8	0.050	0.051	0.053	0.051	0.053	0.051

‘placebo non-responders’ than in ‘drug responders’. In scenario 4 and scenario 6, the effect size in ‘drug responders’ has been increased. In the last three scenarios, the effect size has been set to 0 in either Stage I, or in ‘placebo non-responders’ or ‘drug responders’ in Stage II.

Table 2.3: Power for testing $H_0 : \Delta_1 = 0 \cap \Delta_2 = 0 \cap \Delta_3 = 0$

p_1	q_1	p_2	q_2	p_3	q_3	N	Score Test	GEE $\omega_1=0.5$	GEE $\omega_1=0.6$	GEE $\omega_1=0.7$	GEE $\omega_1=0.8$	GEE $\omega_1=0.9$
0.4	0.3	0.4	0.3	0.9	0.8	412	0.79	0.71	0.76	0.77	0.75	0.68
0.5	0.3	0.4	0.2	0.9	0.8	128	0.80	0.65	0.71	0.76	0.77	0.74
0.5	0.3	0.4	0.1	0.9	0.8	96	0.82	0.64	0.70	0.73	0.71	0.65
0.4	0.3	0.4	0.3	0.9	0.7	312	0.80	0.81	0.82	0.80	0.73	0.61
0.5	0.3	0.4	0.2	0.9	0.7	104	0.81	0.70	0.74	0.75	0.73	0.66
0.5	0.3	0.4	0.1	0.9	0.7	80	0.82	0.66	0.71	0.71	0.67	0.60
0.4	0.4	0.4	0.3	0.9	0.8	2612	0.79	0.96	0.91	0.74	0.43	0.14
0.4	0.3	0.4	0.3	0.9	0.9	728	0.80	0.48	0.62	0.72	0.81	0.83
0.4	0.3	0.4	0.4	0.9	0.8	644	0.80	0.72	0.80	0.84	0.85	0.83

As Table 2.3 shows, $\omega_1 = 0.7$, $\omega_2 = \omega_3 = 0.15$ generally performs the best among the five different combinations of weights in the GEE method, which is consistent with what we observed in analyses for continuous outcomes. However, it should be noted that this

combination of weights also has a relatively large type I error across different response probabilities. Investigators may want to strike a balance between type I error and power when they choose the weights. Additionally, it was noted that when all of the three parts have treatment effect, i.e., $p_1 \neq q_1$ and $p_2 \neq q_2$ and $p_3 \neq q_3$, the score test has a better performance than the GEE method. However, if there is no treatment effect in only one of the three parts, a weight combination can always be found with the GEE method that achieves higher power than the score test. In scenario 7, there is no treatment effect in Stage I. Therefore, the GEE method with larger weights in the second stage gives higher power than the score test. In contrast, in scenario 8 and scenario 9, there is no treatment effect in ‘placebo non-responders’ and ‘drug responders’, respectively. It is not surprising that the GEE method with higher weights in Stage I achieves higher power than the score test.

2.3.2 Continuous outcomes

We also did a broad simulation to assess the performance of three analysis methods for continuous outcomes in the TED. Under the assumption of a normally distributed outcome, for a full specification of the design, all the parameters in Figure 2.2 need to be specified with variance-covariance parameters. Parameters δ_1 , δ_3 , and δ_7 denote the treatment effects in Stage I and Stage II. For example, these parameters are set to zero under the null hypothesis for the evaluation of type I error. As noted in Doros *et al.* [8], parameters δ_0 , δ_{01} , δ_{02} , δ_2 , and δ_4 are the characteristics of subjects who receive only placebo. These parameters can be informed from previous trials. Parameters δ_2 and δ_4 represent the placebo response among ‘placebo non-responders’ and ‘placebo responders’, respectively. Both can be elicited from historical data. Similarly, values for parameters δ_5 and $\delta_6 + \delta_7$ present the characteristics of subjects who receive only the active drug; that is, the drug response among ‘drug non-responders’ and ‘drug responders’, respectively. These parameters can also be elicited from historical data. In this chapter, we assume these parameters for mean changes in Stage II

in proportion to those in Stage I if the subjects receive the same treatment.

Doros *et al.* [8] showed the parameters δ_0 , δ_{01} , and δ_{02} in the placebo group closely connected to the placebo non-response rate. The elicitation of the non-response rate will determine these three parameters. A similar relationship is identified between the non-response rate and δ_{11} and δ_{12} in the active drug group. Let (Y_1, Y_2) represent the outcome at baseline and at the end of Stage I. We assume (Y_1, Y_2) jointly follow a normal distribution, which can be specified separately for the placebo group and the active drug group as follows:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_1 \\ \mu_{21} \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_1 \tau_2 \rho \\ \tau_1 \tau_2 \rho & \tau_2^2 \end{pmatrix} \right]$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_1 \\ \mu_{22} \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_1 \tau_2 \rho \\ \tau_1 \tau_2 \rho & \tau_2^2 \end{pmatrix} \right]$$

Doros *et al.* defined ‘placebo non-responders’ as subjects for whom the outcome at the end of Stage I is both at least half the baseline value and above a certain fixed non-response threshold Ψ . It is understood that the mechanism of drug response might be different from that of placebo response. The definition and extent of ‘responder’ under placebo or the active drug can be completely different. It is worth a discussion between physicians and statisticians during the design phase. Instead, to simplify the situation to facilitate the calculation and simulation, we are using the same definition to identify ‘non-responders’ in the active drug group. The subjects who do not meet one of the criteria above are considered as ‘responders’. As an extension of the work by Doros *et al* [8], we have

$$\begin{aligned}
\mu_{DNR}p_{DNR} &= E(Y_2\{2Y_2 > Y_1 \& Y_2 > \Psi\}) = \int b_2\{2b_2 > b_1, b_2 > \Psi\}f_{Y_1, Y_2}(b_1, b_2)db_1db_2 \\
&= \int_{\Psi}^{\infty} b_2 \int_{-\infty}^{2b_2} f_{Y_1, Y_2}(b_1|b_2)db_1]f_{Y_2}(b_2)db_2 \\
&= \int_{\Psi}^{\infty} b_2\Phi\left(\frac{2b_2 - \mu_1 - \frac{\tau_1}{\tau_2}\rho(b_2 - \mu_{22})}{\tau_1\sqrt{1 - \rho^2}}\right) f_{Y_2}(b_2)db_2 = \\
&= \int_{\frac{\Psi - \mu_{22}}{\tau_2}}^{\infty} (\tau_2 z + \mu_{22})\Phi(az + b)f(z)dz = \tau_2 E\left[Z\Phi(aZ + b)\left\{Z > \frac{\Psi - \mu_{22}}{\tau_2}\right\}\right] + \mu_{22}p_{DNR}
\end{aligned}$$

$$\begin{aligned}
\mu_{DNR} &= \frac{\tau_2}{p_{DNR}} E\left[Z\Phi(aZ + b)\left\{Z > \frac{\Psi - \mu_{22}}{\tau_2}\right\}\right] + \mu_{22} \\
\delta_{11} &= \mu_{DNR} - \mu_{22} = \frac{\tau_2}{p_{DNR}} E\left[Z\Phi(aZ + b)\left\{Z > \frac{\Psi - \mu_{22}}{\tau_2}\right\}\right] \\
p_{DNR}\delta_{11} &= (1 - p_{DNR})\delta_{12}
\end{aligned}$$

where μ_{DNR} is the mean outcome in ‘drug non-responders’, and p_{DNR} is the probability of being a ‘drug non-responder’. Under the alternative hypothesis, μ_{22} is known as $\mu_{21}-\delta_1$. And μ_{DNR} can be simulated from data if ‘non-responder’ is defined as $2Y_2 > Y_1 \& Y_2 > \Psi$. Therefore, we can easily derive δ_{11} and δ_{12} .

With all these parameters and variance-covariance parameters specified, outcomes at baseline, the end of Stage I, and the end of Stage II will be simulated from a multivariate normal distribution. Once data generated, the ‘non-responder’ status will be determined based on the observed data and the prespecified criteria. In our simulation study, we assume the same covariance matrix for the two treatment groups; however, the proposed framework can be easily extended to accommodate different covariance structures. The generated data will be used in both the repeated measures model and the weighted repeated measures model.

In the simulation study, the following parameters were set to be the same as in Doros’

manuscript [8]: the mean and standard deviation of the outcome at baseline were set to 31 and 5, respectively. Standard deviation from baseline to the end of Stage I and to the end of Stage II was set to 7. The correlation between baseline and the changes in both Stage I and Stage II was set to 0.1. Non-response was defined as both change from baseline to the end of Stage I not in excess of half of the baseline outcome value and an outcome value at the end of Stage I greater than 16. The non-response rate was set to 0.75. We assumed mean changes in Stage II to be 75% of the corresponding changes in Stage I for subjects who stay on the same treatment.

We considered the following sample sizes: 100, 120, 160, 200, and 400 subjects with 1:1 randomization at baseline. As we would have four groups: placebo-placebo, placebo-drug, drug-drug, and drug-placebo, each group had 25, 30, 40, 50, and 100 subjects. Correlations between the change in outcome during Stage I and the change in the outcome during Stage II were assumed to be the same for all treatment arms and equal to -0.5, -0.4, -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4, and 0.5. The values of δ_1 , δ_3 , and δ_7 were set to 0 when assessing the type I error. When evaluating the power and mean square error (MSE), δ_1 , δ_3 , and δ_7 were chosen to be 0.3 standard deviation of change from the baseline.

Under these settings, the parameters δ_0 , δ_{01} , and δ_{02} were calculated by using Monte Carlo methods and the formulas developed by Doros *et al.*[8]. Similarly, the parameters δ_{11} and δ_{12} were calculated by using the formulas developed above. Data for baseline, change from baseline to the end of Stage I, and the change from the end of Stage I to the end of Stage II were generated for ‘placebo non-responders’, ‘placebo responders’, ‘drug non-responders’, and ‘drug responders’ from multivariate normal distributions using the mean vector and covariance matrix mentioned above. For each scenario, we generated 10,000 data sets for the evaluation and comparison of the proposed methods. We set the weights for the three parts of the effect size as $\omega_1 = 0.7$, $\omega_2 = 0.15$, and $\omega_3 = 0.15$, to align with other analyses of the TED (Chen *et al.*[5]).

In addition, we also compared the proposed repeated measures model and weighted

repeated measures model with the approach proposed by Chen *et al.*[4]. Three analysis of covariance (ANCOVA) models were constructed by applying Chen’s method to TED. The first ANCOVA model uses all the data from Stage I to estimate δ_1 as the difference in the baseline adjusted mean outcome changes between the two treatment groups, whereas the second and third ANCOVA model uses the data on Stage II ‘placebo non-responders’ and ‘drug responders’ to estimate δ_3 and δ_7 as the difference in baseline (the outcome value at the end of Stage I) adjusted mean outcome changes between the two treatment groups. It was shown that the estimates δ_1 , δ_3 , and δ_7 are not correlated. Therefore, under the null hypothesis, the test statistic is written as

$$T_{\text{OLS}} = \frac{\omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_3 + \omega_3 \hat{\delta}_7}{\sqrt{\omega_1^2 \text{Var}(\hat{\delta}_1) + \omega_2^2 \text{Var}(\hat{\delta}_3) + \omega_3^2 \text{Var}(\hat{\delta}_7)}}$$

Type I Error As shown in Figure 2.3, for almost all the sample sizes, the repeated measures model has the lowest type I error. It is well controlled below or around 0.05. This is also correct across all the correlation values. We also examined the performance of the weighted repeated measures model. When the sample size increases, the weighted repeated measures model, with weights from K-means clustering, performs continuously better under the scheme of this approach, especially for large K. However, the variability of the weighted repeated measures model with weights from propensity scores is higher than the other settings. We also noticed that even though the performance is improved with the increase of the sample size, the type I error for the weighted repeated measures model is slightly inflated, no matter how the weights are estimated. However, the investigator can always tune the calibration parameter K to get the type I error appropriately controlled by using the weights from the K-means clustering approach. Chen’s OLS method has an average performance between that of the repeated measures model and that of the weighted repeated measures model.

MSE of Treatment Effect Compared with the other methods, the weighted repeated measures model with weights from propensity scores performs very well for all the sample

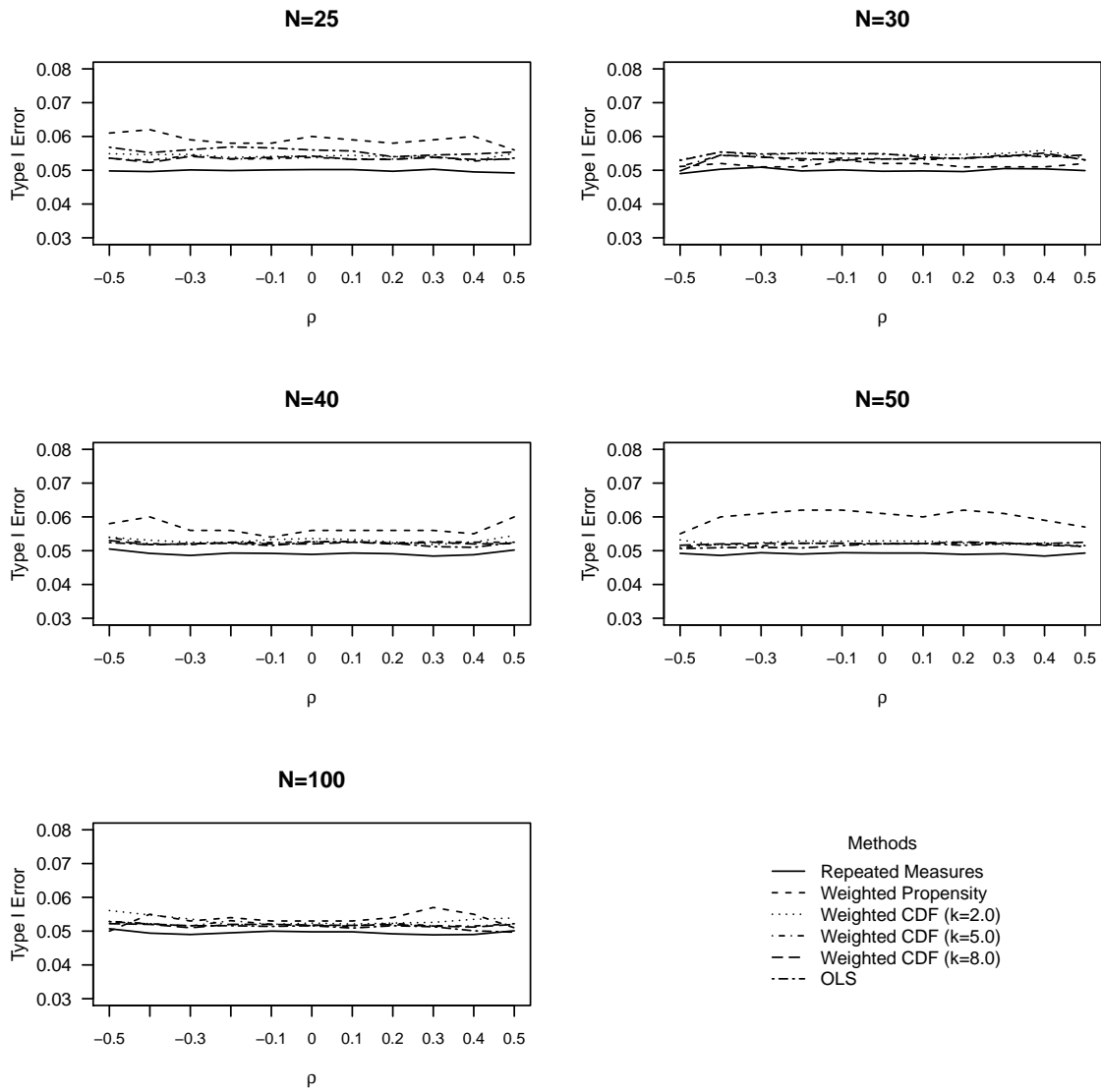


Figure 2-3: Type I error within the framework of TED

sizes and different correlation values. The K-means-weighted repeated measures models have similar performance as the propensity score weighted repeated measures model. In addition, the repeated measures model has a consistently small MSE across all the different models and settings (Figure 2-4). It is noteworthy that, as the sample size increases fourfold from 25 subjects to 100 subjects per group, the MSE reduces to one-fourth of the original value,

from around 1.4 to 0.35.

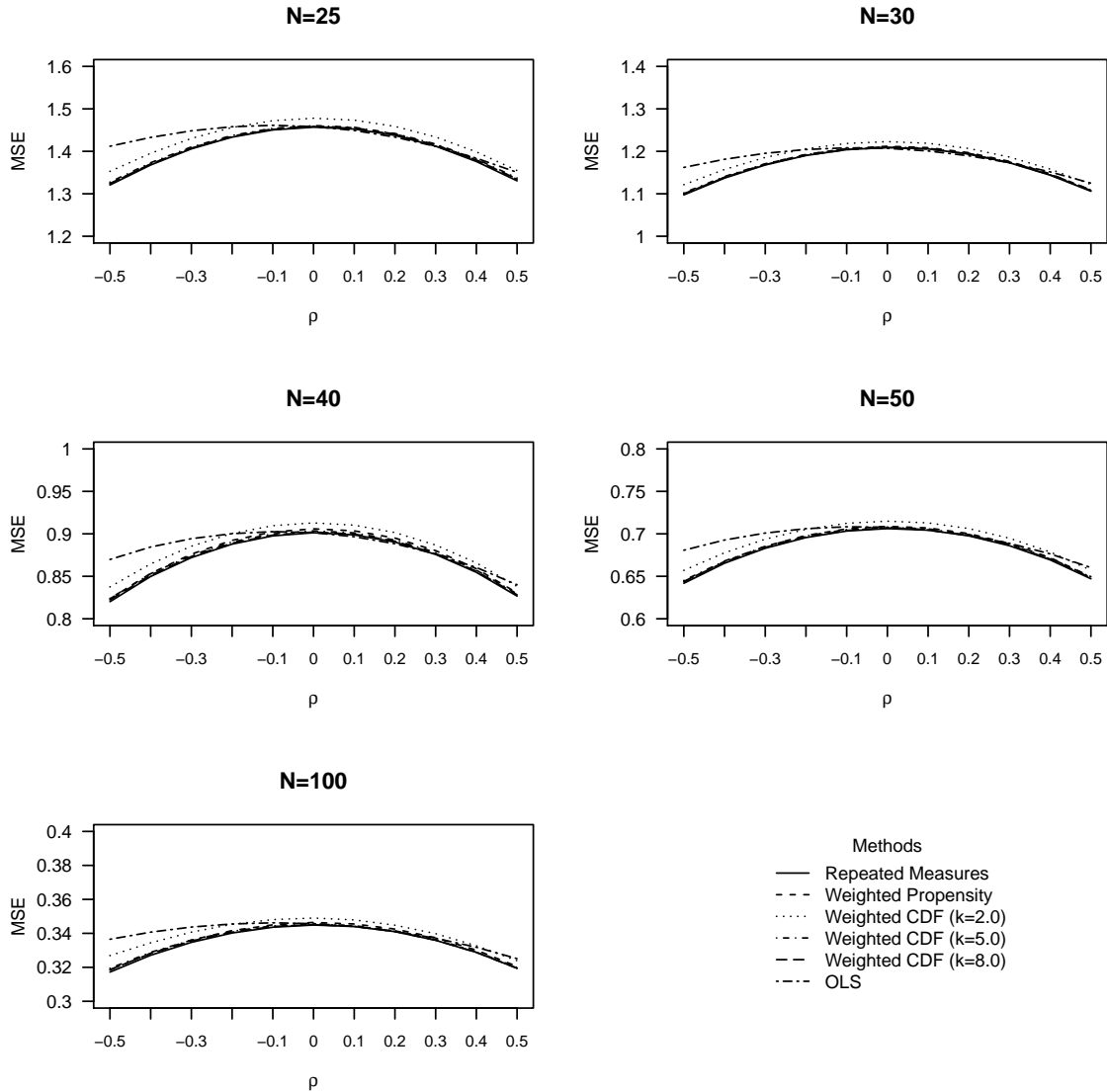


Figure 2-4: MSE within the framework of TED

Power Figure 2-5 compares the power across different models. The repeated measures model achieves the highest power. The weighted repeated measures model with weights from propensity scores has the lowest power for all sample sizes and across different correlation values. But it is notable that there is no significant difference in power between these methods. The investigators need only consider type I error and MSE when they want to

choose from the repeated measures model and the weighted repeated measures model. Of note, we saw no big differences between the OLS method and the proposed approaches.

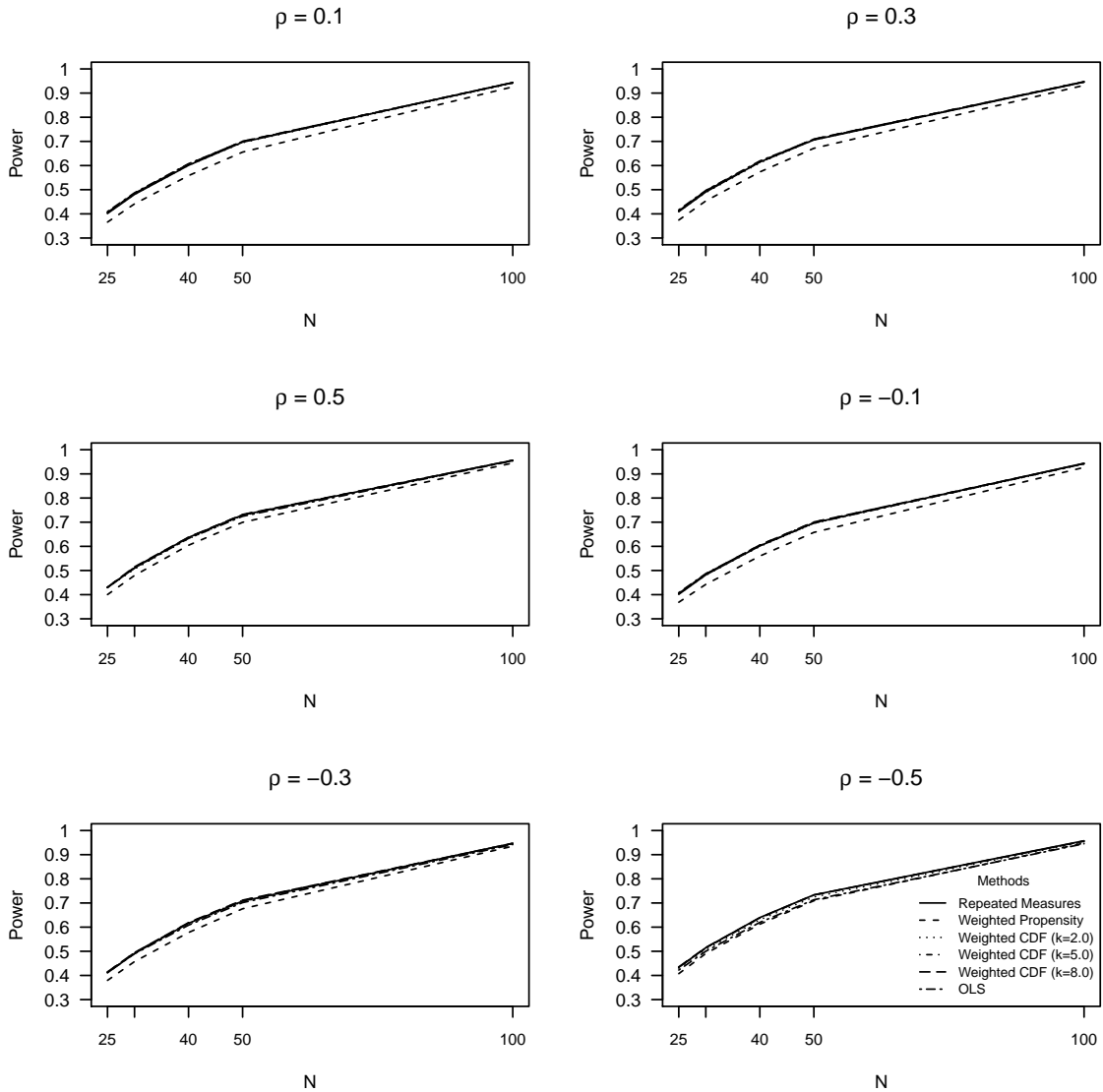


Figure 2-5: Power within the framework of TED

It is encouraging to see consistent results across the three metrics: type I error, MSE of treatment effect, and power. The repeated measures model is the obvious first choice if not much prior information is available about the treatment effect. It gives a robust estimate of treatment effect, provides a well-controlled test, and achieves higher power if there is

no difference between the two groups. If there is some expectation for the effect size, the weighted repeated measures model is an alternative option. It uses more data from the second stage and provides investigators with an opportunity to adjust the distribution of the weights.

2.4 Conclusions

In all sequential designs with enrichment strategies, the main target is to reduce the placebo response in the estimate of treatment effect. SPCD tries to separate ‘placebo responders’ in the first stage and to involve ‘placebo non-responders’ in the second stage mainly. The TED takes one more step to also rule out ‘drug non-responders’ in the first stage on the basis that the treatment effect is very small or not detectable in ‘placebo responders’ and ‘drug non-responders’. However, many considerations should be taken when applying the TED in practice. As Ivanova *et al.* note, TED requires that the active drug has a significantly superior performance compared with placebo in achieving short-term efficacy. The trial might lose a large portion of subjects if the treatment duration is long, as half of the ‘drug responders’ will be assigned to placebo. All efforts should be taken to minimize the duration of the two stages in order to limit the dropout of subjects. When TED was first proposed, it was only available for trials with binary outcomes. However, analysis methods were needed to address continuous outcomes in psychiatry research. In this chapter, we proposed a new method for assessing binary outcomes and several new models for assessing continuous outcomes in TED trials. All of these methods use all the data collected in the trial to estimate the treatment effect.

The repeated measures model can be seen as an alternative approach to the available score tests when the outcome is binary. When there is no difference in treatment effect in only one of the three components, it has been shown that the repeated measures model obtains a higher power to detect the overall treatment effect. At the stage of trial design, if the investigators believe there is no treatment effect in Stage I, more weights can then be put

on the data collected in the second stage; otherwise, investigators can locate more weights at Stage I. Investigators can adjust the weights to achieve the optimal power based on the scenario in practice. This approach provides investigators with more flexibility during the trial design phase. However, it also requires the investigators to have a good estimation of the effect size in the whole population, including the ‘placebo non-responders’ and the ‘drug responders’. It will be a challenge for the investigators to determine the appropriate weights for the first-stage data and for ‘placebo non-responders’ and ‘drug responders’ in the second stage. Investigators can choose the weights based on prior data from existing trials, or they can follow the suggested weights ($\omega_1 = 0.7$, $\omega_2 = \omega_3 = 0.15$) in this chapter. Although it is not the scope of this chapter to discuss the selection of weights, it is worth further exploration.

For TED trials with continuous outcomes, the repeated measures model and the weighted repeated measures model may be used for the analysis of the overall treatment effect. Both methods treat the outcomes from the two stages as repeated measures and take the correlations within each individual into account when estimating the overall treatment effect. The only difference between these two models is in the description of the placebo response. The repeated measures model uses a binary placebo response status with an absolute cut-off in the outcome and a percent change from baseline outcome. The weighted repeated measures model describes placebo non-response as a pre-specified characteristic, to some degree, in every subject. This measure of placebo non-response was proposed to be used as a weight in the analysis in the second stage in the weighted repeated measures model. From our observation of the imputation studies, if the investigators have prior information about the relationship between placebo response and subject characteristics, the weighted repeated measures model will be a good choice, as it uses more measurements to evaluate a subject’s probability of being a ‘placebo non-responder’. If there is no prior information, the investigators can still use the repeated measures model to analyze the trial data.

Chapter 3

Treatment Response as Latent Binary Characteristics in Two-way Enriched Design

3.1 Background and Motivation

As mentioned in the previous chapter, TED was initially designed to work for binary outcomes [14]. Ivanova *et al.* proposed a variety of score tests with one, two, or three degrees of freedom, depending on different possible assumptions one is willing to make about the underlying parameters. However, in many clinical trials, the outcome of interest is a continuous variable. Liu *et al.* [19] extended the scope of TED to continuous outcomes. They illustrated the feasibility by the implementation of different methods to the analysis of continuous outcomes within the TED framework. These methods include the use of an OLS approach [4], a repeated measures model [8], and a weighted repeated measures model [29]. A major difference in these methods is how to analyze the correlation between outcomes in Stage I and Stage II. The OLS method assumes a strong constancy of correlation between stages, which leads to the independence of the estimates of treatment effect in Stage I and Stage II. However, the other two models attempt to describe the correlation between stages.

In the framework of TED, it generally assumes subjects are either ‘responders’ or ‘non-responders’ [14, 15, 37], according to their responses at the end of Stage I based on the pre-specified criteria. Both the OLS and the repeated measures model consider placebo response as a binary status. However, such classification might be misleading, due to measurement error and uncertainty of the criteria. In order to reduce the impact of misclassification of the subjects, the weighted repeated measures model views the placebo response as a continuous measure that is present to some degree in each subject. Placebo response is handled as a weight instead of a dichotomous variable in the weighted repeated measures model [19]. The weighting allows all the subjects in the placebo group in Stage I enter the second stage and contributes to the estimate of the treatment effect in Stage II according to their measures of placebo non-response.

However, all these current methods consider the response status as a measurable variable. They all try to classify subjects into ‘responders’ or ‘non-responders’ based on a pre-specified criterion. The classification is based on the observed outcomes at the end

of Stage I. However, such criterion-based approaches have several limitations, as noted by Fava *et al.* [9] and Rybin *et al.* [30]. First of all, the cutoff point is usually selected based on clinical judgment and experience from previous studies. Subjects with outcomes around the cutoff point can be classified into totally different subgroups even though there is a small difference among them. Rybin *et al.* [30] evaluated the extent and consequences of misclassification through a simulation study with data from a recent SPCD trial [10]. They reported the impact of change in the criterion cutoff on the estimation of the treatment effect in SPCD. By using the same trial data, a 10% change in cutoff from 50% to 40% resulted in one standard error change in the SPCD treatment effect. In addition, measurement error could have a high impact on criterion-based methods. It was noted by Fava *et al.* [9] that the measurement error can be one of the causes of the observed placebo response. All the previous analyses and evaluations reveal the fact that the classification of subjects based on a single criterion may lead to great bias in the estimate of treatment effect. TED needs a more comprehensive and robust analysis approach.

In this chapter, we propose to view the response status to the active drug or placebo as a characteristic of each subject. Subjects are either ‘placebo responders’ or ‘placebo non-responders’ by nature, which can only be revealed if they are in the placebo group. Similarly, subjects are either ‘drug responders’ or ‘drug non-responders’ by nature, which can only be revealed if they are in the active drug group. However, such characteristics cannot be observed or measured accurately in a clinical trial. Rybin *et al.* [30] proposed to consider placebo response as a latent variable in the SPCD framework. It is reflected and estimated in the placebo group. In our work, we extend this idea to both treatment groups, so that whether a subject responds to the active drug or placebo are both considered as latent variables. We observe the outcome in all subjects without knowing the actual placebo response status nor drug response status of each subject. Thus, the sample distribution of the observed outcome can be viewed as two parts: one part as an unlabeled mixture of the outcome in ‘placebo responders’ and ‘placebo non-responders’ in the placebo group

and modeled as a mixture of two distributions with the placebo response probability as the mixing probability; and the other part as an unlabeled mixture of the outcome in ‘drug responders’ and ‘drug non-responders’ in the active drug group and modeled as a mixture of two distributions with the drug response probability as the mixing probability. The placebo response probability can be estimated in the placebo group in Stage I; similarly, the drug response probability can be estimated in the drug group in Stage I.

Many methods have been proposed for estimating parameters in the mixture of normal distribution. These approaches were summarized in a comprehensive review by Redner and Walker [27]. Dempster *et al.* [6] shown the expectation-maximization (EM) algorithm is superior to the other methods despite some limitations, including slow convergence. We will extend Rybin’s work [30] and implement the EM algorithm to estimate the parameters of the mixture of normal distributions in the scheme of TED. The estimate of the mixing probability in this manuscript is based on the likelihood of response. There is no distributional assumption on the analysis.

Several analysis approaches have been applied to the analysis of continuous outcomes in the TED. They include the OLS approach, the repeated measures model, and the weighted repeated measures model. These three methods are considered as criterion-based analysis approaches as they all use pre-specified criteria to determine subjects’ response status. We will compare the proposed approach with these methods to evaluate the performance of the new method. The OLS approach has a strong assumption and considers independence between outcomes in Stage I and Stage II. It also considers the placebo response as a binary status. The repeated measures model relaxes the independence assumption, but still considers the placebo response as a dichotomous variable. The weighted repeated measures model includes more flexibility by removing the independence assumption and introducing uncertainty to the placebo response. The proposed approach moves forward one more step by considering the placebo response and drug response as latent characteristics and introducing stochastic components in the estimation of the treatment effect. We believe

that the comparison of these four methods covers the spectrum of assumptions of outcomes and the approach to placebo response and drug response.

This chapter is organized as follows. In the first part, we describe the problem parameterizations, specify the likelihood, provide the estimate of parameters and variance components, and define the treatment effect within the proposed scheme. In the second part, we present the simulation results, comparisons between the proposed method and other existing approaches, and an example of the application of this proposed method to ADAPT-A study data. Finally, we provide our extensive discussion and conclusion.

3.2 Methodology

3.2.1 Response as a latent characteristic

Suppose we analyze data from a TED trial with two stages of equal length. In addition, let us assume subjects can be accurately classified into ‘placebo responders’ and ‘placebo non-responders’ if they are treated with placebo, and can be correctly classified into ‘drug responders’ and ‘drug non-responders’ if they are under the treatment of the active drug. Figure 3-1 shows a graphical presentation of the effects in the TED study with respect to the true placebo response and drug response characteristics.

‘Placebo responders’ and ‘drug responders’ are shown in light grey and dark grey rectangles. ‘Non-responders’ are shown in white squares. The dash lines represent outcome progression in different subgroups: δ_1 and δ_2 are changes in ‘placebo non-responders’ and ‘placebo responders’ in Stage I placebo group; δ_3 and δ_4 are changes in ‘drug non-responders’ and ‘drug responders’ in Stage I drug group; δ_{11} and δ_{12} are changes in Stage II in ‘placebo non-responders’ randomized to placebo and the active drug; δ_{41} and δ_{42} are changes in Stage II in ‘drug responders’ randomized to placebo and the active drug. If the placebo response probability is π_1 and the drug response probability is π_2 , the Stage I treatment effect in the overall sample is $\delta_I = (\pi_2\delta_4 + (1 - \pi_2)\delta_3) - (\pi_1\delta_2 + (1 - \pi_1)\delta_1)$, the Stage II treatment effect in ‘placebo non-responders’ is $\delta_{NR}^p = \delta_{12} - \delta_{11}$, and the Stage II treatment effect in ‘drug

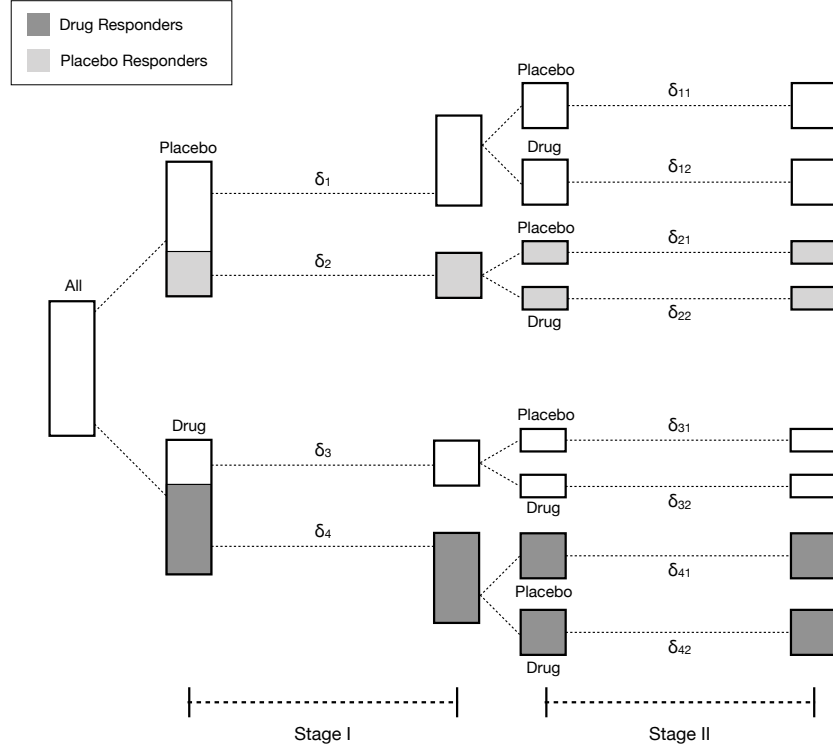


Figure 3.1: Two-way enriched design

responders' is $\delta_R^d = \delta_{42} - \delta_{41}$. Therefore, for a specified set of TED weights ω_1 , ω_2 , and ω_3 , the TED treatment effect is defined as the linear combination $\delta_\omega = \omega_1 \delta_I + \omega_2 \delta_{NR}^p + \omega_3 \delta_R^d$.

Suppose the severity of the disease is measured three times over the study: at baseline (Y_{01}), at the end of Stage I (Y_{02}), and at the end of Stage II (Y_{03}). The outcome for Stage I is defined as the change in the severity of the disease from baseline to the end of Stage I ($Y_1 = Y_{02} - Y_{01}$), and the outcome for Stage II is defined as the change in the severity of the disease from the end of Stage I to the end of Stage II ($Y_2 = Y_{03} - Y_{02}$).

Let g_{1i} and g_{2i} be binary indicators for the Stage I and Stage II group assignment (placebo = 0, drug = 1) for subject i . Let $R_i^p = 0$ or 1 be an indicator of being a 'placebo responder' and $R_i^d = 0$ or 1 be an indicator of being a 'drug responder' for subject i

depending on which treatment group the subject is assigned to. The R_i^p and R_i^d are two unobserved (latent) characteristics that are assumed to follow binomial distribution $R_i^p \sim B(1, \pi_{1i})$ and $R_i^d \sim B(1, \pi_{2i})$, where π_{1i} is the probability of being a ‘placebo responder’ and π_{2i} is the probability of being a ‘drug responder’.

Table 3.1: TED parametrization: Distributions

Treatment Group	Stage I	Stage II
Placebo Non-Responders	$p_{101}(Y_1 Y_{01}) \sim N(\mu_{101}, \sigma_{101}^2)$	$p_{201}(Y_2 Y_1) \sim N(\mu_{201}, \sigma_{201}^2)$
Placebo Responders	$p_{102}(Y_1 Y_{01}) \sim N(\mu_{102}, \sigma_{102}^2)$	$p_{202}(Y_2 Y_1) \sim N(\mu_{202}, \sigma_{202}^2)$
Drug Non-Responders	$p_{103}(Y_1 Y_{01}) \sim N(\mu_{103}, \sigma_{103}^2)$	$p_{203}(Y_2 Y_1) \sim N(\mu_{203}, \sigma_{203}^2)$
Drug Responders	$p_{104}(Y_1 Y_{01}) \sim N(\mu_{104}, \sigma_{104}^2)$	$p_{204}(Y_2 Y_1) \sim N(\mu_{204}, \sigma_{204}^2)$

Under the assumption of normality, the outcome distributions in the TED study groups can be summarized as shown in Table 3.1. The means at the end of Stage I can be modeled as a function of baseline and group assignment. The means at the end of Stage II can be modeled as a function of change in Stage I and group assignment. The formulas are shown in Table 3.2. Here, b_{22} and b_{28} are Stage II treatment effects in ‘placebo non-responders’ and ‘drug responders’, correspondingly.

Table 3.2: TED parametrization: Mean functions

Treatment Group	Stage I	Stage II
Placebo Non-Responders	$\mu_{101} = b_{01} + b_{11}Y_{01}$	$\mu_{201} = b_{02} + b_{12}Y_1 + b_{22}g_2$
Placebo Responders	$\mu_{102} = b_{03} + b_{13}Y_{01}$	$\mu_{202} = b_{04} + b_{14}Y_1 + b_{24}g_2$
Drug Non-Responders	$\mu_{103} = b_{05} + b_{15}Y_{01}$	$\mu_{203} = b_{06} + b_{16}Y_1 + b_{26}g_2$
Drug Responders	$\mu_{104} = b_{07} + b_{17}Y_{01}$	$\mu_{204} = b_{08} + b_{18}Y_1 + b_{28}g_2$

3.2.2 Parameter estimation with EM algorithm

The joint distribution of Y_1 and Y_2 for Stage I placebo subjects can be presented as a Gaussian mixture of ‘placebo responder’ and ‘placebo non-responder’ distributions with mixture parameters π_1 . Under the same consideration, the joint distribution of Y_1 and Y_2 for Stage I drug subjects can be presented as a Gaussian mixture of ‘drug responder’ and ‘drug non-responder’ distributions with mixture parameters π_2 .

Stage I placebo group:

$$p(y_{1i}, y_{2i}, R_i^p) = p(R_i^p)p(y_{1i}, y_{2i}|R_i^p) = \pi_1^{R_i^p} (1 - \pi_1)^{1-R_i^p} p(y_{1i}, y_{2i}|R_i^p)$$

$$p(y_{1i}, y_{2i}|R_i^p = 0) = p_{101}(y_{1i})p_{201}(y_{2i}|y_{1i})$$

$$p(y_{1i}, y_{2i}|R_i^p = 1) = p_{102}(y_{1i})p_{202}(y_{2i}|y_{1i})$$

Stage I active drug group:

$$p(y_{1i}, y_{2i}, R_i^d) = p(R_i^d)p(y_{1i}, y_{2i}|R_i^d) = \pi_2^{R_i^d} (1 - \pi_2)^{1-R_i^d} p(y_{1i}, y_{2i}|R_i^d)$$

$$p(y_{1i}, y_{2i}|R_i^d = 0) = p_{103}(y_{1i})p_{203}(y_{2i}|y_{1i})$$

$$p(y_{1i}, y_{2i}|R_i^d = 1) = p_{104}(y_{1i})p_{204}(y_{2i}|y_{1i})$$

Therefore, the full likelihood for total sample size N can be written as follows:

$$L = \prod_{i=1}^N [(1 - \pi_2)^{(1-R_i^d)} \pi_2^{R_i^d} [p_{103}(y_{1i})p_{203}(y_{2i}|y_{1i})]^{(1-R_i^d)} [p_{104}(y_{1i})p_{204}(y_{2i}|y_{1i})]^{R_i^d}]^{g_{1i}}$$

$$[(1 - \pi_1)^{(1-R_i^p)} \pi_1^{R_i^p} [p_{101}(y_{1i})p_{201}(y_{2i}|y_{1i})]^{(1-R_i^p)} [p_{102}(y_{1i})p_{202}(y_{2i}|y_{1i})]^{R_i^p}]^{(1-g_{1i})}$$

The expected value of response for i -th individual R_i^p distributed as $B(1, \pi_{1i})$ is π_{1i} , and R_i^d distributed as $B(1, \pi_{2i})$ is π_{2i} . Therefore, after the integration of log likelihood, the Q-function takes the following form:

$$\begin{aligned}
Q &= \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})[\log p_{101}(y_{1i}) + \log p_{201}(y_{2i}|y_{1i})] \\
&+ \sum_{i=1}^N (1 - g_{1i})\pi_{1i}[\log p_{102}(y_{1i}) + \log p_{202}(y_{2i}|y_{1i})] \\
&+ \sum_{i=1}^N g_{1i}(1 - \pi_{2i})[\log p_{103}(y_{1i}) + \log p_{203}(y_{2i}|y_{1i})] \\
&+ \sum_{i=1}^N g_{1i}\pi_{2i}[\log p_{104}(y_{1i}) + \log p_{204}(y_{2i}|y_{1i})]
\end{aligned}$$

In general, the response probability for i -th subject at iteration $s+1$ can be calculated from the mixture probability function and current estimates of the parameters $\theta^{(s)}$ and the placebo response $\pi_1^{(s)}$ and the drug response $\pi_2^{(s)}$.

$$\pi_{1i}^{(s+1)} = \frac{\pi_1^{(s)} p_{102}(y_{1i}|\theta^{(s)}) p_{202}(y_{2i}|y_{1i}, \theta^{(s)})}{\pi_1^{(s)} p_{102}(y_{1i}|\theta^{(s)}) p_{202}(y_{2i}|y_{1i}, \theta^{(s)}) + (1 - \pi_1^{(s)}) p_{101}(y_{1i}|\theta^{(s)}) p_{201}(y_{2i}|y_{1i}, \theta^{(s)})}$$

$$\pi_{2i}^{(s+1)} = \frac{\pi_2^{(s)} p_{104}(y_{1i}|\theta^{(s)}) p_{204}(y_{2i}|y_{1i}, \theta^{(s)})}{\pi_2^{(s)} p_{104}(y_{1i}|\theta^{(s)}) p_{204}(y_{2i}|y_{1i}, \theta^{(s)}) + (1 - \pi_2^{(s)}) p_{103}(y_{1i}|\theta^{(s)}) p_{203}(y_{2i}|y_{1i}, \theta^{(s)})}$$

Here we define the placebo response as a function of the outcome change in Stage I alone in placebo group. It is determined only through the outcome collected in Stage I. Similar consideration is applied to the drug response. Therefore, we can reconstruct the formulas as,

$$\begin{aligned}
\pi_{1i}^{(s+1)} &= \frac{\pi_1^{(s)} p_{102}(y_{1i}|\theta^{(s)})}{\pi_1^{(s)} p_{102}(y_{1i}|\theta^{(s)}) + (1 - \pi_1^{(s)}) p_{101}(y_{1i}|\theta^{(s)})} \\
\pi_{2i}^{(s+1)} &= \frac{\pi_2^{(s)} p_{104}(y_{1i}|\theta^{(s)})}{\pi_2^{(s)} p_{104}(y_{1i}|\theta^{(s)}) + (1 - \pi_2^{(s)}) p_{103}(y_{1i}|\theta^{(s)})}
\end{aligned}$$

Then, the placebo response probability π_1 and the drug response probability π_2 are calculated as the average of individual placebo response probabilities π_{1i} and the average of individual drug response probabilities π_{2i} , respectively.

$$\pi_1^{(s+1)} = \frac{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i}^{(s+1)}}{\sum_{i=1}^N (1 - g_{1i})}$$

$$\pi_2^{(s+1)} = \frac{\sum_{i=1}^N g_{1i} \pi_{2i}^{(s+1)}}{\sum_{i=1}^N g_{1i}}$$

At each iteration of the EM algorithm, the model parameters are estimated. All distribution parameters can be maximized explicitly using Q-function derivatives. Appendix A presents the formulations for all model parameters.

The covariance matrix of the parameters can be estimated at each iteration by the inverse of the observed Fisher information matrix $\Sigma = I^{-1}(\hat{\theta})$, where $I(\hat{\theta}) = -(\frac{\partial^2 Q(\theta|Y_o)}{\partial \theta^2})_{\theta=\hat{\theta}}$, θ is the vector of parameters, and Y_o is the observed outcome data. The elements of the Hessian matrix are listed in Appendix B.

3.2.3 The definition of treatment effect

The Stage I treatment effect is constructed from parameters and their variances estimated by EM. The Stage I change in the drug group is constructed as a mixture of mean changes in ‘drug responders’ and ‘drug non-responders’.

$$\hat{\mu}_{11} = \pi_2 \hat{\mu}_{104} + (1 - \pi_2) \hat{\mu}_{103} = \pi_2 \left(\hat{b}_{07} + \hat{b}_{17} \frac{\sum_{i=1}^N g_{1i} y_{01i}}{\sum_{i=1}^N g_{1i}} \right) + (1 - \pi_2) \left(\hat{b}_{05} + \hat{b}_{15} \frac{\sum_{i=1}^N g_{1i} y_{01i}}{\sum_{i=1}^N g_{1i}} \right)$$

The Stage I change in the placebo group is constructed as a mixture of mean changes in ‘placebo responders’ and ‘placebo non-responders’.

$$\hat{\mu}_{10} = \pi_1 \hat{\mu}_{102} + (1 - \pi_1) \hat{\mu}_{101} = \pi_1 \left(\hat{b}_{03} + \hat{b}_{13} \frac{\sum_{i=1}^N (1 - g_{1i}) y_{01i}}{\sum_{i=1}^N (1 - g_{1i})} \right) + (1 - \pi_1) \left(\hat{b}_{01} + \hat{b}_{11} \frac{\sum_{i=1}^N (1 - g_{1i}) y_{01i}}{\sum_{i=1}^N (1 - g_{1i})} \right)$$

The variance of each mixture component $\hat{\sigma}_{101}^2$, $\hat{\sigma}_{102}^2$, $\hat{\sigma}_{103}^2$, and $\hat{\sigma}_{104}^2$ is a sum of the elements of the corresponding covariance matrix. The variance of the Stage I change in the active drug group is calculated as follows.

$$\hat{\sigma}_{11}^2 = \pi_2^2 \hat{\sigma}_{104}^2 + (1 - \pi_2)^2 \hat{\sigma}_{103}^2$$

The variance of the Stage I change in the placebo group is calculated as follows.

$$\hat{\sigma}_{10}^2 = \pi_1^2 \hat{\sigma}_{102}^2 + (1 - \pi_1)^2 \hat{\sigma}_{101}^2$$

Therefore, Stage I treatment effect is $\hat{\delta}_I = \hat{\mu}_{11} - \hat{\mu}_{10}$ with the corresponding variance $\text{Var}(\hat{\delta}_I) = \hat{\sigma}_{11}^2 + \hat{\sigma}_{10}^2$.

The Stage II treatment effect in ‘placebo non-responders’ and ‘drug responders’ and their variances are directly estimated with the EM algorithm as \hat{b}_{22} , \hat{b}_{28} , $\text{Var}(\hat{b}_{22})$, and $\text{Var}(\hat{b}_{28})$.

The TED effect weights ω_1 , ω_2 , and ω_3 are specified. A test for $H_0 : \delta_\omega = \omega_1 \delta_I + \omega_2 \delta_{NR}^p + \omega_3 \delta_R^d = 0$ is based on the test statistic

$$T = \frac{\omega_1 \hat{\delta}_I + \omega_2 \hat{\delta}_{NR}^p + \omega_3 \hat{\delta}_R^d}{\sqrt{\omega_1^2 \text{Var}(\hat{\delta}_I) + \omega_2^2 \text{Var}(\hat{\delta}_{NR}^p) + \omega_3^2 \text{Var}(\hat{\delta}_R^d)}}$$

We assume T to follow approximately the standard normal distribution under the null hypothesis.

3.3 Simulation Study

3.3.1 Parameter setting

A broad simulation was undertaken to evaluate the performance of the proposed EM method against the three known methods. Data were generated from multivariate distributions, for the four patient subgroups: ‘placebo non-responders’, ‘placebo responders’, ‘drug non-responders’, and ‘drug responders’. Rybin *et al.* [30] provided a good summary of consider-

ations in the data simulation for the evaluation of the analytical methods. The first point is about the connection between the non-response rate and outcomes. As illustrated in Doros *et al.* [8], the parameters of outcomes in the placebo group are closely connected to the placebo non-response rate. A similar relationship in the active drug group is identified by Liu *et al.* [19]. The second point is related to how much difference in the mean outcomes between ‘responders’ and ‘non-responders’. This directly impacts the performance of the analytical methods in the identification of ‘responders’ and ‘non-responders’ in the placebo group and the active drug group.

Rybin *et al.* [30] illustrated the relationship between non-response rate and the Stage I changes in the placebo group. A similar correlation in the active drug group is identified and presented below. Suppose that in drug subjects the distribution of outcome at baseline (Y_{01}) and the end of Stage I (Y_{02}) is a mixture of multivariate normal distributions with equal covariance and mixing probability π_2 . We assume the covariance matrix is the same in the ‘drug responders’ and ‘drug non-responders’. However, the same covariance matrix constraint can be removed without difficulty. The proposed framework can be easily extended to accommodate different covariance structures.

$$p(Y_{01}, Y_{02}) = (1 - \pi_2)p_3(Y_{01}, Y_{02}) + \pi_2p_4(Y_{01}, Y_{02})$$

$$p_3(Y_{01}, Y_{02}) \sim N \left[(\mu_{013}, \mu_{023}), \begin{pmatrix} \sigma_{01}^2 & \rho\sigma_{01}\sigma_{02} \\ \rho\sigma_{01}\sigma_{02} & \sigma_{02}^2 \end{pmatrix} \right]$$

$$p_4(Y_{01}, Y_{02}) \sim N \left[(\mu_{014}, \mu_{024}), \begin{pmatrix} \sigma_{01}^2 & \rho\sigma_{01}\sigma_{02} \\ \rho\sigma_{01}\sigma_{02} & \sigma_{02}^2 \end{pmatrix} \right]$$

It is understood that the mechanism of drug response might be different from that of placebo response. The definition and extent of response under placebo or under the active drug can be completely different. It is worth a discussion between physicians and statisticians during the design phase. Instead, to simplify the situation and to facilitate the calculation and simulation, we will use the same definition for drug non-response as

for placebo non-response and define drug non-response in terms of parameter Ψ_1 and Ψ_2 and both $\Psi_1 Y_{02} > Y_{01}$ (improvement during Stage I) and $Y_{02} > \Psi_2$ (severity at the end of Stage I). These criteria will be used by the three criterion-based analysis methods to determine subjects' response status. However, this assumption can be adjusted accordingly to fit various clinical trials. The probability of non-response can be presented as follows.

$$\begin{aligned}
p_{DNR} &= P(\Psi_1 Y_{02} > Y_{01}, Y_{02} > \Psi_2) \\
&= \int \int_{y_{01} < \Psi_1 y_{02}, y_{02} > \Psi_2} \left[(1 - \pi_2) p_3(y_{01}, y_{02}) + \pi_2 p_4(y_{01}, y_{02}) \right] dy_{01} dy_{02} \\
&= (1 - \pi_2) \int_{\Psi_2}^{\infty} \left[\int_{-\infty}^{\Psi_1 y_{02}} p_3(y_{01} | y_{02}) dy_{01} \right] p_3(y_{02}) dy_{02} \\
&\quad + \pi_2 \int_{\Psi_2}^{\infty} \left[\int_{-\infty}^{\Psi_1 y_{02}} p_4(y_{01} | y_{02}) dy_{01} \right] p_4(y_{02}) dy_{02} \\
&= (1 - \pi_2) \int_{\frac{\Psi_2 - \mu_{023}}{\sigma_{02}}}^{\infty} \Phi(a_3 z + b_3) f(z) dz + \pi_2 \int_{\frac{\Psi_2 - \mu_{024}}{\sigma_{02}}}^{\infty} \Phi(a_4 z + b_4) f(z) dz \\
&= (1 - \pi_2) E \left[\Phi(a_3 z + b_3) | Z > \frac{\Psi_2 - \mu_{023}}{\sigma_{02}} \right] + \pi_2 E \left[\Phi(a_4 z + b_4) | Z > \frac{\Psi_2 - \mu_{024}}{\sigma_{02}} \right]
\end{aligned}$$

where $a_3 = a_4 = \frac{\Psi_1 \sigma_{02} - \sigma_{01} \rho}{\sigma_{01} \sqrt{1 - \rho^2}}$, $b_3 = \frac{\Psi_1 \mu_{023} - \mu_{013}}{\sigma_{01} \sqrt{1 - \rho^2}}$, $b_4 = \frac{\Psi_1 \mu_{024} - \mu_{014}}{\sigma_{01} \sqrt{1 - \rho^2}}$, and Z is a standard normal variable, and f and Φ are the density and cumulative distribution function of a standard normal distribution. This relationship indicates that for a given set of parameters in 'drug non-responders' and drug response rate ($1 - p_{DNR}$), we can calculate the Stage I change in 'drug responders'. The calculation can be performed with Monte Carlo simulation.

The data were simulated from a multivariate Normal distribution. The following parameters were set to be the same as in Rybin's manuscript [30]. The mean and standard deviation of the outcome at baseline for all four groups were set to 31 and 5, respectively. The probability of response to placebo and the active drug were both set to 0.3 ($\pi_1 = \pi_2 = 0.3$). Criterion-based non-response was defined as both change from baseline

to the end of Stage I not in excess of half of the baseline outcome value ($\Psi_1 = 2$) and an outcome value at the end of Stage I greater than 16 ($\Psi_2 = 16$). The mean change from baseline to the end of Stage I in ‘placebo responders’ was set to 16. The mean change from baseline to the end of Stage I in ‘placebo non-responders’ can be derived with the formula provided by Rybin *et al.* [30]. The mean change from baseline to the end of Stage I in ‘drug responders’ can be derived with the formula presented above by setting the mean change from baseline to the end of Stage I in ‘drug non-responders’. However, in the simulation study, it will be based on the null hypothesis or the alternative hypothesis settings. Under the null hypothesis, we set the mean change from baseline to the end of Stage I in ‘drug responders’ and ‘drug non-responders’ the same as those in ‘placebo responders’ and ‘placebo non-responders’, respectively. Under the alternative hypothesis, supposing the treatment effect in Stage I (δ_I) equals one, we set the mean change from baseline to the end of Stage I in ‘drug non-responders’ to 11, which led to the mean change from baseline to the end of Stage I in ‘drug responders’ to 16.1, according to the definition of treatment effect in Stage I $\delta_I = (\pi_2\delta_4 + (1 - \pi_2)\delta_3) - (\pi_1\delta_2 + (1 - \pi_1)\delta_1)$. This setting ensures that the outcome changes in drug groups are not smaller than those in the placebo groups, and are much close to the relationship between non-response rate and the Stage I changes in the drug group, as shown above. Standard deviations of Stage I and Stage II changes were set to 2 in Stage I placebo group and set to 5 in Stage I drug group. The correlation between baseline and the changes in Stage I and Stage II was set to 0.1. Mean changes in Stage II were assumed to be 40% smaller than the corresponding changes in Stage I in subjects who remained on the same treatment from Stage I to Stage II. A summary of the fixed parameters under the alternative hypothesis ($\delta_I=1$) is presented in Table 3.3.

We considered the following sample sizes: 200, 300, 400, 600, and 800 subjects, with Stage I randomization 1:1 in the placebo group and the active drug group. Correlations between the change in the outcome during Stage I and the change in the outcome during Stage II was assumed to be the same for all treatment arms. The values of δ_I , δ_{NR}^p , δ_R^p , δ_{NR}^d ,

Table 3.3: Simulation parameters

Treatment Group	Y_{01}		Y_1		Y_2	
	Mean	SD	Mean	SD	Mean	SD
Placebo Non-responders						
PP	31	5	9.6	2	0.6×9.6	2
PD	31	5	9.6	2	$0.6 \times 9.6 + \delta_{NR}^p$	2
Placebo Responders						
PP	31	5	16	2	0.6×16	2
PD	31	5	16	2	$0.6 \times 16 + \delta_R^p$	2
Drug Non-responders						
DP	31	5	11	5	$0.6 \times 11 - \delta_{NR}^d$	5
DD	31	5	11	5	0.6×11	5
Drug Responders						
DP	31	5	16.1	5	$0.6 \times 16.1 - \delta_R^d$	5
DD	31	5	16.1	5	0.6×16.1	5

Y_{01} : Baseline outcome.

Y_1 : Outcome change from baseline to the end of Stage I.

Y_2 : Outcome change from the end of Stage I to the end of Stage II.

and δ_R^d were set to 0 under the null hypothesis to evaluate type I error. The parameters are set to 1 under the alternative hypothesis for MSE and power calculations.

For each scenario, we generated 10,000 datasets to ensure enough evaluable datasets left for evaluation of the proposed EM method and the other three criterion-based methods. The weights ω_1 , ω_2 , and ω_3 were set to 0.7, 0.15, and 0.15, respectively, to be consistent with previous analyses. Simulations were run in R 3.6.0 and SAS 9.4 on a Linux cluster.

3.3.2 Overall performance

EM algorithm has shown superior performance than other methods in the estimate of parameters of the mixture of normal distribution [27]. However, it has the limitation of slow convergence [6]. In order to get enough evaluable datasets for methods comparison, we

allowed a maximum of 10,000 iterations of the algorithm. In practice, if the calculation doesn't converge in the limited number of iterations, we recommend the investigators to take the average of the estimates from the last 100 iterations as the final estimates.

The estimated proportions of 'placebo responders' (placebo response probability) and 'drug responders' (drug response probability) under the alternative hypothesis are presented in Table 3.4. The true response probability was set to (0.3, 0.3) for the pair of placebo response probability and drug response probability. As expected, with the increase of sample size, the bias of the estimates decreases. The accuracy of the estimate increases as a function of available information. As also noticed in Rybin *et al.* [29], the degree of correlation does not have a big impact on the estimate of response probability. The estimated probability of response moves closer to the true values, with approximately the same rate for all correlation values. Therefore, we only present the results under correlation $\rho = 0.1$. As the formula in Section 3.2.2 presents, the estimate of response probability is highly correlated with the distribution of the outcome. We keep the standard deviation of the outcome change from baseline to the end of Stage I (Y_1) in placebo group unchanged ($SD_1 = 2$) and reduce the standard deviation of Y_1 in the active drug group (SD_2) from 5 to 2. The change in the estimate of placebo response probability is trivial in these scenarios as the distribution of the outcome in the placebo group does not change. However, in the active drug group, large sample size is needed to get a less biased estimate of drug response probability for the cases with big standard deviations than those with small standard deviations. The accuracy of the estimate is increasing with the decrease of variability in the outcome.

3.3.3 Comparison with other methods

We compared the EM method to the other three methods, the OLS method, the repeated measures model, and the weighted repeated measures model, with the simulated datasets. Several scenarios were under consideration. In the first scenario, it was assumed that the response threshold was specified correctly for the methods other than the EM method,

Table 3.4: Estimates of the true response probabilities (0.3, 0.3) with $\rho = 0.1$

	N=50	N=75	N=100	N=150	N=200
$SD_2 = 5$	(0.381, 0.514)	(0.363, 0.501)	(0.349, 0.490)	(0.334, 0.482)	(0.323, 0.467)
$SD_2 = 4$	(0.383, 0.446)	(0.363, 0.433)	(0.340, 0.424)	(0.334, 0.405)	(0.323, 0.385)
$SD_2 = 3$	(0.384, 0.375)	(0.364, 0.357)	(0.351, 0.350)	(0.335, 0.342)	(0.324, 0.334)
$SD_2 = 2$	(0.385, 0.309)	(0.364, 0.304)	(0.352, 0.301)	(0.335, 0.300)	(0.324, 0.300)

and the treatment effects were equal in ‘responders’ and ‘non-responders’. This is the primary analysis for the performance evaluation. In order to test the robustness of the EM method, we also performed additional analyses by changing different parameters. We first relaxed the assumption of equal treatment effects, but still kept the assumption of correctly specified response threshold. Then, we assessed how the EM method performed under the misspecification of threshold but with equal treatment effects. Furthermore, we relaxed both assumptions and studied the combined influence of threshold misspecification and inequality of treatment effects on the performance of the methods. TED assumes a high proportion of ‘placebo responders’ in the population. In the fourth scenario, we evaluated the robustness of the EM method when there was no ‘placebo responder’ in the population. Additionally, we explored the impact of different proportions of outcome changes from Stage I to Stage II in the active drug group and the placebo group, and the impact of various standard deviations of outcome changes in Stage I in the active drug group.

Primary Analysis: Correct response threshold, equal treatment effects in ‘responders’ and ‘non-responders’

Type I Error Figure 3.2 presents the type I error under different parameter settings. In the majority of the various cases, these four methods have the type I error under 0.05. The EM method doesn’t achieve the minimum type I error, but it still controls the type I

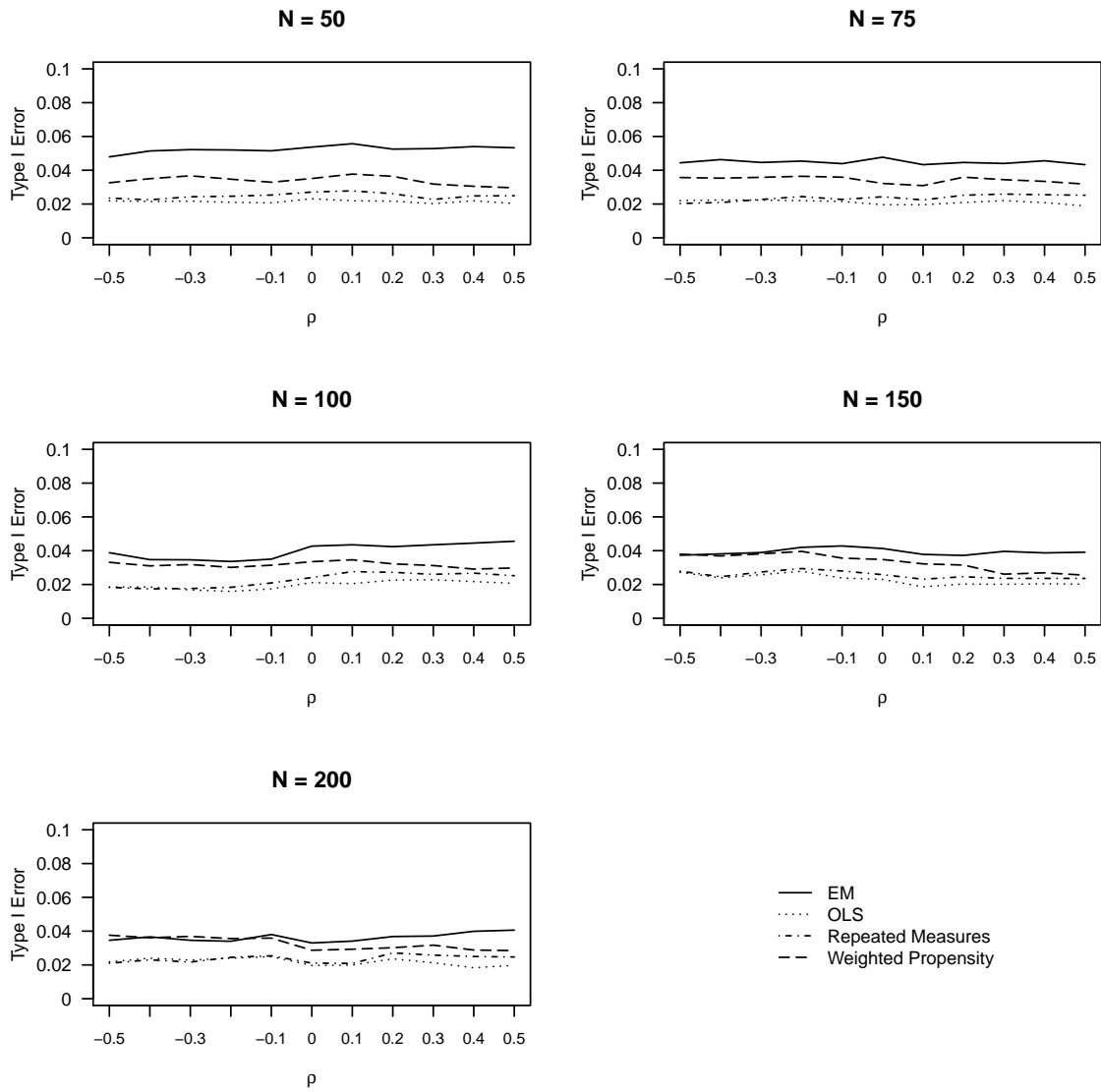


Figure 3-2: Type I error assuming correct response threshold and equal treatment effects

error around or just below 0.05. With the increase of the sample size, the type I error is decreasing, with different rates in these four methods. The reduction is evident in the EM method. It indicates the proposed EM algorithm requires more sample size than the other methods to have the type I error controlled. This is a potential challenge for this approach to be applied in reality.

MSE Figure 3.3 compares the MSE of the treatment effect in different methods for the simulated data. The MSE has a similar level across the different methods, except in the weighted repeated measures model. The weighted repeated measures model requires two logistic regressions to get pseudo weights for both the placebo group and the active drug group. It might be more difficult for this method to get a less biased estimate. The EM method has the minimum MSE in different parameter settings. However, the difference in MSE across the different methods is trivial.

Power Power is presented in Figure 3.4. It is great to see that the EM method is superior to the other methods in almost all the parameter settings. When the correlation between the outcome changes in Stage I and Stage II is positive, the repeated measures model also has an outstanding performance. It even has higher power than the EM algorithm when the sample size is large.

Robustness Analysis 1: Correct response threshold, different treatment effects in ‘responders’ and ‘non-responders’

We evaluated the performances of the four methods when the treatment effect was different in ‘responders’ and ‘non-responders’. We kept the treatment effect in ‘non-responders’ δ_{NR} unchanged but varied the treatment effect in ‘responders’ δ_R . Specifically, we changed the value of h in the function $\delta_R = h\delta_{NR}$. When h is between 0 and 1, then the treatment effect in ‘responders’ is smaller than that in ‘non-responders’. When h is larger than 1, then the treatment effect in ‘responders’ is larger than that in ‘non-responders’. The special case is that when h is 0, then there is no treatment effect in ‘responders’. In Figure C.1, MSE is evaluated when the treatment effect in ‘responders’ is different from that of ‘non-responders’. In general, the weighted repeated measures model has larger variation in the MSE with the change in the value of h than the other three methods. The OLS method and the repeated measures model have similar MSE values as the EM method, but the EM method has the smallest variation in the MSE. In Figure C.3, Figure C.5, and Figure

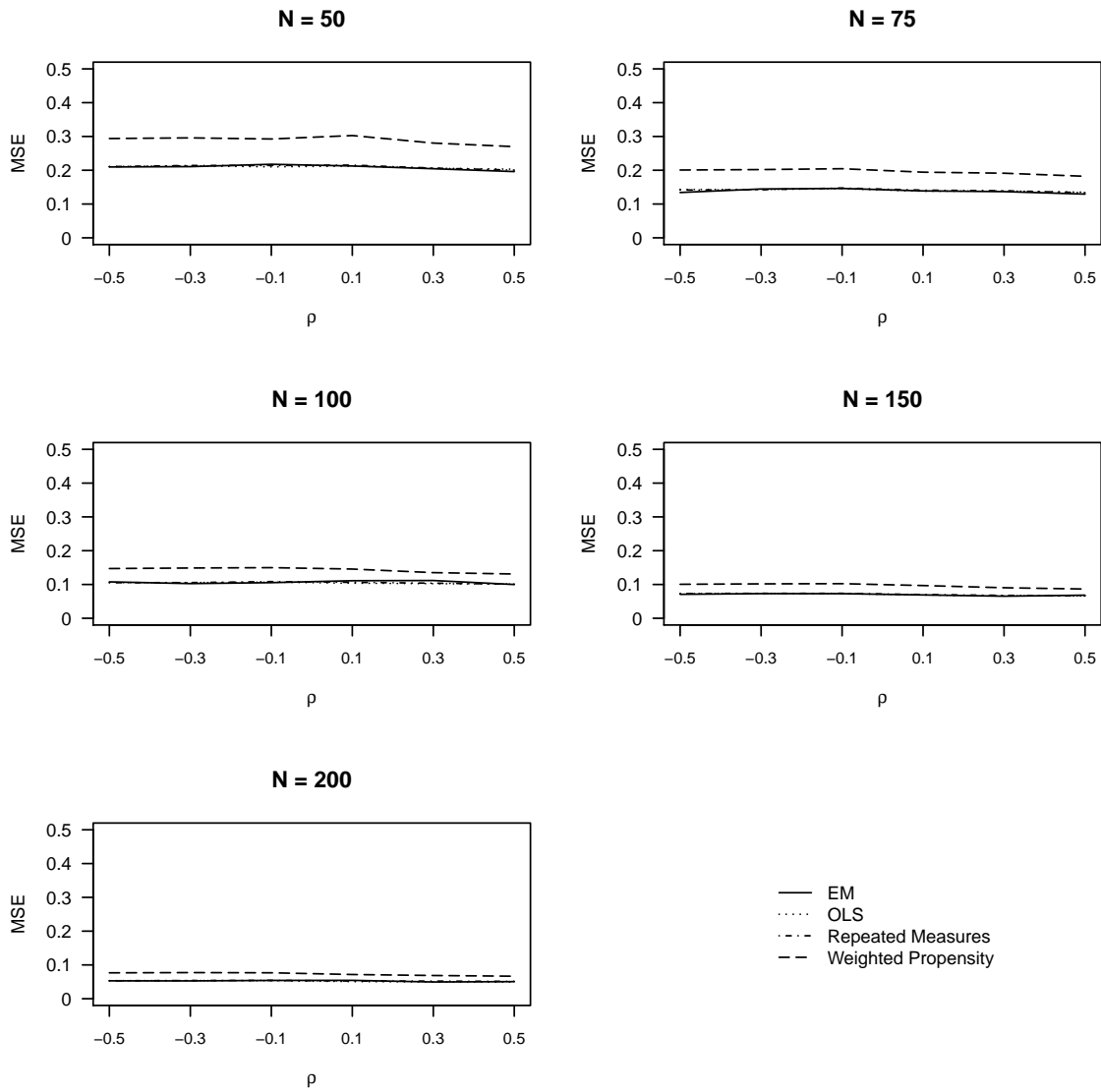


Figure 3-3: MSE assuming correct response threshold and equal treatment effects

C-7, power is evaluated when the treatment effect in ‘responders’ is set to half of that in ‘non-responders’, two times of that in ‘non-responders’, or zero. The corresponding MSE in these three cases are presented in Figure C-2, Figure C-4, and Figure C-6. The proposed EM method has a consistently good performance in all of these parameter settings. The repeated measures model can reach a higher power than the EM method with large sample

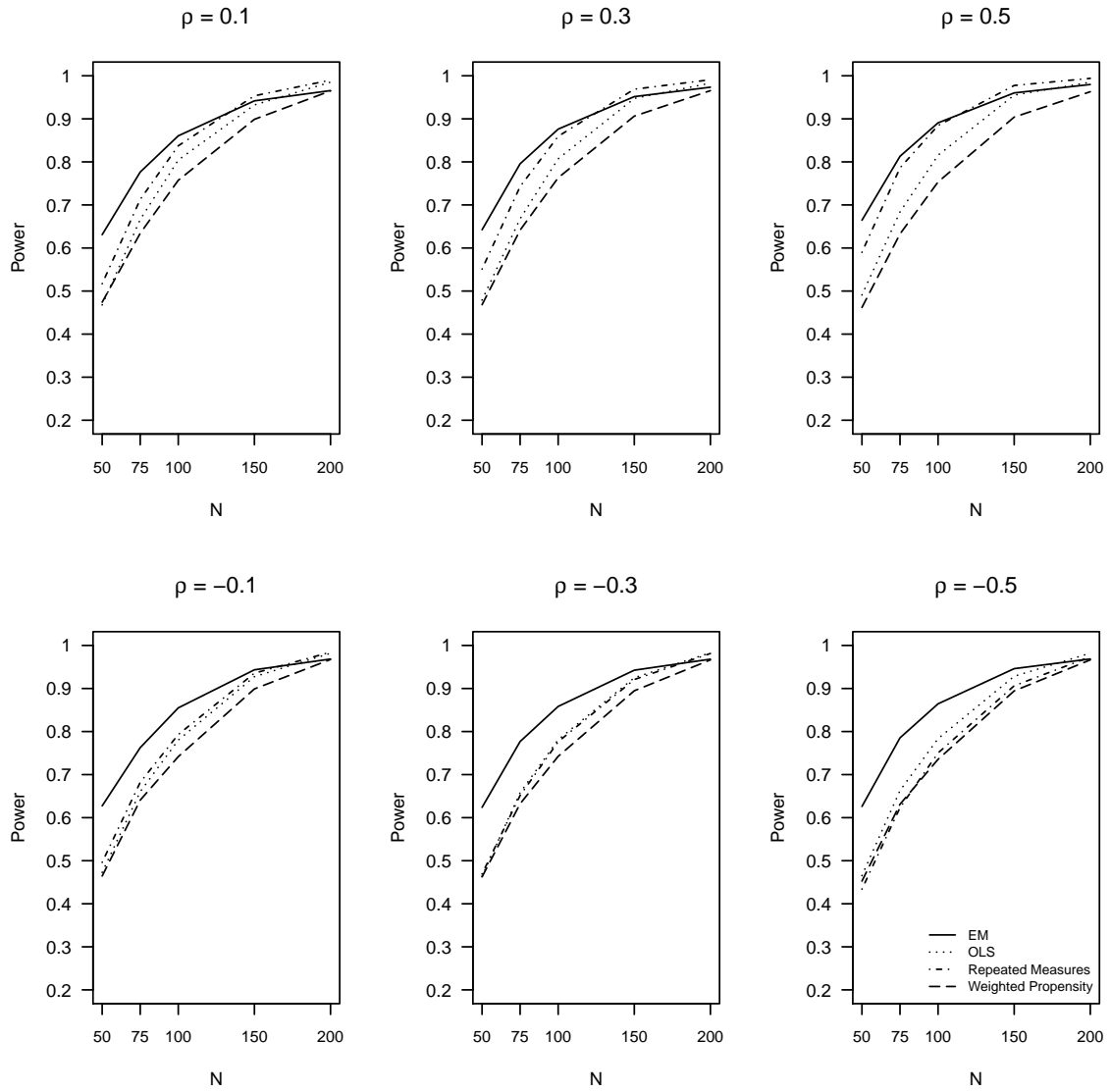


Figure 3-4: Power assuming correct response threshold and equal treatment effects

size when the correlation is positive. However, it is sensitive to the parameter setting.

Robustness Analysis 2: Response threshold misspecification, equal treatment effect in ‘responders’ and ‘non-responders’

We assessed the performance of the proposed method when the response threshold is mis-

specified. In data generation, we used a threshold of 60% to simulate the sample data. But in data analysis, we still used 50% as the threshold to determine whether a subject was a ‘responder’ or not. In order to evaluate the pure influence of response threshold misspecification, we restricted equal treatment effects in ‘responders’ and ‘non-responders’. Figure C-8 and Figure C-9 compare the methods in terms of MSE and power. We omitted the OLS method due to its poor performance in this scenario.

Table 3.5: Departure in Stage I outcome changes in the active drug group

	δ_1	δ_2	δ_3	δ_4	Departure $ \delta_1 - \delta_3 + \delta_2 - \delta_4 $
Case 1 (Primary Analysis)	9.6	16	11	16.1	$1.4 + 0.1 = 1.5$
Case 2 (Robustness Analysis)	9.6	16	9.6	19.3	$0 + 3.1 = 3.1$
Case 3 (Robustness Analysis)	9.6	16	8	23.1	$1.6 + 7.1 = 8.7$

As mentioned in Section 3.3.1, in the simulation study, the Stage I outcome changes in the active drug group were selected according to the definition of the overall treatment effect in Stage I $\delta_I = (\pi_2\delta_4 + (1 - \pi_2)\delta_3) - (\pi_1\delta_2 + (1 - \pi_1)\delta_1)$, which did not align with the relationship between the outcome change and the non-response rate in Stage I in the drug group required by the definition of non-response for the criterion-based methods. However, the two types of repeated measures models and the OLS method will still use the pre-specified criteria to determine the subjects’ response status. Data generation might impact the performance of these three criterion-based methods. We considered the outcome changes in the placebo group as the standards and defined the departure as the sum of absolute differences between the placebo group and the active drug group. Table 3.5 presents the three cases in consideration and evaluation. From the primary analysis, we further decreased the outcome changes in Stage I in the ‘drug non-responders’, which led to an increase in the corresponding changes in ‘drug responders’. This change further deviates from the relationship between the outcome change and the non-response rate in Stage I in

the drug group. The performance of the three criterion-based methods is getting worse as the deviation is more and more severe. However, the EM algorithm seems not sensitive to this deviation. Figure C-18 and Figure C-19 present the power in this scenario.

Robustness Analysis 3: Response threshold misspecification, different treatment effects in ‘responders’ and ‘non-responders’

In this scenario, we relaxed both restrictions and evaluated the influence of response threshold misspecification and different treatment effects. The treatment effect in ‘responders’ was set to half of that in ‘non-responders’. The thresholds in simulation and data analysis were still set to 60% and 50%, respectively. Figure C-10 and Figure C-11 summarize the comparison among the four methods on MSE and power. We omitted the OLS method due to its poor performance in this scenario.

Robustness Analysis 4: No ‘placebo responder’

The TED and the proposed EM algorithm are trying to reduce the impact of placebo response, which makes the assumption that the expected proportion of ‘placebo responders’ is high enough to warrant the concern for a possible bias of treatment effect estimate [30]. In addition to the three scenarios mentioned above, we also tested the fourth scenario that there is no ‘placebo responder’ in the population. Figure C-12 and Figure C-13 illustrate the MSE and power when there is no ‘placebo responder’ in the population.

Robustness Analysis 5: Variation in the standard deviation of the outcome change in Stage I

We changed the standard deviation of the outcome change in Stage I in the drug group from 5 to 7. As the data are more scattered, it is more difficult for the EM algorithm to get a controlled type I error, especially for a trial with small sample size. This trend can be easily seen in Figure C-14 and Figure C-15.

Robustness Analysis 6: Different proportions in outcome changes from Stage I to Stage II in the active drug group and the placebo group

In the primary analysis, we assumed that the outcome change in Stage II is 60% of that in Stage I in those who remained on the same treatment in the two stages for both groups. In this scenario, we changed this percentage to 80% for subjects who took the active drug in both stages. Type I error is shown in Figure C-16. It is easy to notice that the type I error gets larger for the two repeated measures models when the correlation between the outcome changes in the two stages turns to positive. This inflation is more obvious with the increase of the correlation. A similar difference can be observed in power, as shown in Figure C-17. As the TED uses Stage II data from both the placebo group and drug group and assigns the same weights on these two components, the increase of proportion in the drug group breaks this balance. In addition, the positive correlation has the same direction as the proportion, which enlarges the departure from the assumptions. The two types of repeated measures models are sensitive to this deviation and respond to it with dramatic changes. Therefore, the performance of the repeated measures model and the weighted repeated measures model are not satisfactory when the correlation is positive.

In all these robustness analyses, the proposed EM method has a better performance than the other three methods in almost all the parameter settings. The robustness of the proposed EM method provides the investigators with an option when the assumption of equal treatment effects or response threshold specification is unclear.

3.3.4 Antidepressant therapy trial example

The ADAPT-A study is a multi-center, double-blind, placebo-controlled study of the efficacy of low-dose aripiprazole (2 mg/day) adjunctive to antidepressant therapy (ADT) in the treatment of major depressive disorder patients with a history of inadequate response to

prior antidepressant treatment [10]. It was originally conducted under the SPCD framework. In this 60-day trial with two equal-length stages, subjects were initially randomized in a 2:3:3 ratio to treatment sequences of aripiprazole/aripiprazole, placebo/aripiprazole, and placebo/placebo. The primary endpoint was the change from baseline in Montgomery-Åsberg Depression Rating Scale (MADRS) score. Non-response was assessed at the end of Stage I and defined as a less than 50% decrease in MADRS score from baseline and a MADRS score of greater than 16.

For subjects in the placebo group at Stage I, we kept their original outcomes at Stage I and Stage II. We imputed the outcomes for Stage I drug group to reach a 1:1 ratio sample size in the placebo group and the drug group. We randomly assigned treatment group to the subjects who were classified as ‘drug responders’ at the end of Stage I. In the final data, we obtained 167 subjects initially assigned to placebo and 162 subjects to aripiprazole at Stage I. We applied the proposed method to this imputed dataset in the framework of TED. The TED effect δ_ω was calculated as a linear combination of Stage I effect Δ_I in all subjects and Stage II effect Δ_R^d in ‘drug responders’ and δ_{NR}^p in ‘placebo non-responders’, with corresponding weights 0.7, 0.15, and 0.15. The algorithm converged after 374 iterations. The grouping of 167 Stage I placebo subjects and 162 Stage I drug subjects based on their Stage I progression are different from the criterion-based method. The EM method estimated mean placebo response rate at 39%, compared to a rate of 22% with the criterion-based method (Figure C·21). The EM method estimated mean drug response rate at 37%, compared to a rate of 30% with criterion-based methods (Figure C·22).

The trial data were assessed by using the OLS method, the repeated measures model, the weighted repeated measures model with propensity score as weights, and the proposed EM method. The treatment effect estimates were compared based on these approaches and summarized in Table 3.6. The convergence plot of the estimates of the treatment effects and response probabilities is presented in Figure C·20.

Table 3.6: Treatment effect estimates for ADAPT-A trial

Method	Stage I Effect	Stage II Effect		TED Effect
		Drug Responders	Placebo Non-responders	
OLS	-5.75 [-7.07, -4.43]	-0.42 [-2.54, 1.70]	-2.17 [-4.38, 0.04]	-4.42 [-5.45, -3.38]
Repeated Measures	-6.28 [-7.53, -5.02]	-0.57 [-2.35, 1.22]	-2.67 [-4.86, -0.48]	-4.88 [-5.85, -3.91]
Weighted Repeated	-7.08 [-8.39, -5.77]	-0.59 [-2.45, 1.27]	-2.71 [-4.92, -0.49]	-5.45 [-6.46, -4.45]
EM Algorithm	-5.77 [-6.84, -4.70]	0.34 [-2.13, 2.81]	-2.35 [-5.28, 0.57]	-4.34 [-5.28, -3.40]

Estimate [95% CI] are presented.

3.4 Conclusions

The treatment effect in TED is a composite of three parts: the treatment effect in the whole population in Stage I, and the treatment effect in the ‘placebo non-responders’ and ‘drug responders’ in Stage II. It is critical to find a robust method to classify subjects based on their responses to the active drug and placebo, as subjects from two subgroups will be used twice in the estimate. The previous methods based on a pre-specified criterion, such as the OLS method, the repeated measures model, and the weighted repeated measures model, may include classification error and lead to a biased inference. The uncertainty of placebo response and drug response should be taken into account in the treatment effect estimation.

In this chapter, we propose to view placebo response and drug response as latent variables. Subjects are either ‘placebo responders’ or ‘placebo non-responders’. Such a characteristic is revealed if they are in the placebo group. On the other hand, subjects are either ‘drug responders’ or ‘drug non-responders’. The drug response characteristic is revealed if they are under the treatment of the active drug. Therefore, the population is a mixture of ‘responders’ and ‘non-responders’. The mixing probability depends on the treatment they are taking. The goal of the analysis is to estimate the treatment effect as a combination of the parameters of the mixture.

Under the assumption of normality, the outcomes are modeled with Gaussian mixture. The likelihood is derived and the EM algorithm is applied for the estimation of response

probabilities and the treatment effect. This newly proposed method is compared with the three methods that are mentioned above. These three competitors cover the spectrum of the assumptions of the outcomes and the association between stages. The proposed method achieves consistent performance in different scenarios. Under the perfect situation of equal treatment and correct classification, the EM algorithm preserves the type I error and has the minimum MSE and high power. When the two constraints are relaxed, the EM algorithm largely reduces the MSE when the treatment effects are different in ‘responders’ and ‘non-responders’; and it reaches remarkably high power than the other three approaches when the threshold of the criterion is misspecified. In addition, it shows a consistent performance when there is no ‘placebo responder’ in the analysis population.

The robust properties of the EM algorithm provide much more confidence to the investigators during the design phase. This approach is extremely useful, given an ambiguous definition of clinical response. It also allows the investigators to have different response probabilities for placebo and the active drug. As Fava *et al.* [9] noted, the degree of placebo response can vary depending on the outcome measure used and on the degree of expectations about the outcome. It will be beneficial for the investigators if the method involves a pragmatic process in the estimation of treatment effect.

However, the EM algorithm is not perfect without limitation. First of all, it takes a longer time to obtain a converged result. Sometimes when it doesn’t converge, we have to take the suboptimal estimates from the last iteration. Secondly, it requires a large sample size to get a less biased estimate of response probability. However, such a requirement will be a cost-constraint for pharmaceutical companies in the conduct of clinical trials. In addition, clear separation of the means of the mixture improves the performance of the EM algorithm, which indicates EM algorithm might not be a wise choice if the means are similar in ‘responders’ and ‘non-responders’.

This chapter introduces a new view of placebo response and drug response in TED as inherit unobserved qualities. It provides a robust approach to the investigators when the

definition of response is unclear. This latent characteristic approach can be successfully used for joint modeling of the observed outcome data and the latent response information. Flexibility is also warranted on varied treatment effects, deviation in classification criterion, and different response probabilities.

Chapter 4

Assessment of the Performance of Sequential Enriched Design

4.1 Background and Motivation

When the underlying patient population consists of a sizable proportion of patients who are either ‘placebo responders’ or who are categorically never ‘responders’, the existing enrichment designs intended to address the problem of high placebo response in clinical trials, such as SPCD and TED, might produce biased estimates of the treatment effect for the target patient population that responds to an active drug but not to placebo. In an attempt to improve the estimation of the target treatment effect for clinical trials that include a mixed patient population, Chen *et al.* [5] proposed SED aiming to exclude subjects who are ‘placebo responders’ and those who never respond to the active drug or placebo from the analysis.

SED is designed to locate the target population, patients who respond to the active drug but not to placebo, through the enrichment process, by combining the idea of placebo lead-in design, SPCD, and TED. Different from SPCD and TED, this design has three stages. The name SED suggests the enrichment is sequential. As shown in Figure 4-1, all subjects are treated with placebo in the first stage (Stage 0 - placebo lead-in stage). At the beginning of the second stage (Stage I), ‘placebo non-responders’ are selected and randomized to placebo or the active drug, while ‘placebo responders’ will switch to the active drug for the following stages. At the beginning of the third stage (Stage II), ‘drug responders’ are further selected based on pre-specified criteria and re-randomized to placebo or the active drug; while ‘drug non-responders’ will go on with the active drug to the end of the trial. The primary feature of SED is to screen out not only patients with high placebo response but also patients who do not respond to the active drug from the study and the analysis.

As the only three-stage design, SED has yet to be implemented in practice. Concerns might come from different perspectives: the complexity of the design, the longer duration of trials, the large sample size required, and the lack of flexible analysis methods. It is stated that the additional stage can help to purify the target population in the analysis

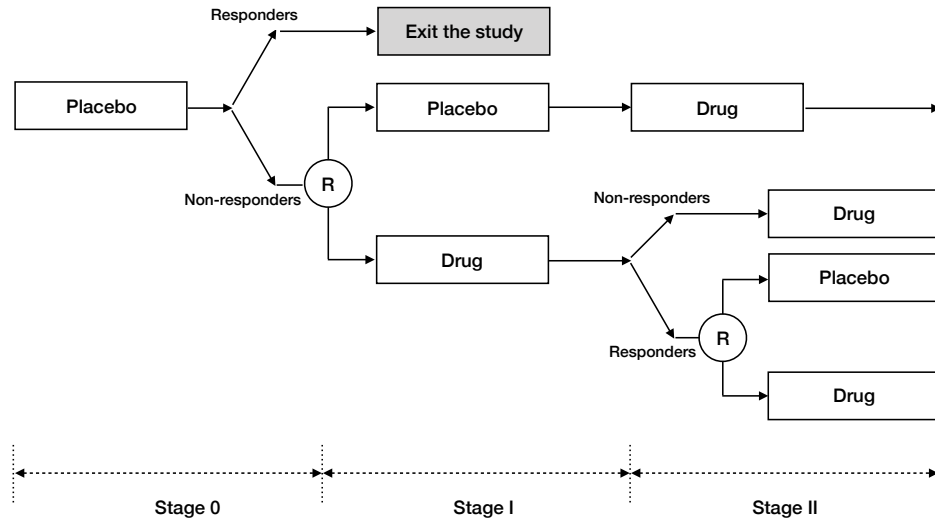


Figure 4.1: Sequential enriched design

sample, and eventually get a less biased estimate of the treatment efficacy. This intention is visible, as we can see from the performance improvement from the one-stage design, parallel randomized design, to the several two-stage designs, first placebo lead-in design, then SPCD and TED. MSE of the estimate of drug efficacy has been largely reduced through the two-stage designs. Power also increases tremendously with the same amount of patients. There is no doubt that keeping adding stages will get a more purified subset. However, the question is, how much we can gain from it and whether we can afford the corresponding cost. The extra stage in SED requires additional subjects at enrollment and extended trial duration. These requirements can create huge financial burdens on pharmaceutical companies, especially in Phase II and Phase III studies. Therefore, critical appraisal is urgently needed to determine whether this three-stage design has substantial advantages over the two-stage designs (placebo lead-in design, SPCD, and TED).

In this chapter, we will first describe the components of the analysis population and parameter setting. The potential concerns in the previous evaluation will be discussed.

These will be the basis of data generation and design comparison. Then, we will perform a simulation study and comprehensively review the requirements and performance of SED in different scenarios. We will also explore the possibility of the application of the newly proposed analysis methods to SED. Finally, we will conclude with our considerations and suggestions.

4.2 Methodology

4.2.1 Description of population and parameter setting

For any two-arm randomized clinical trial, the enrolled population can always be classified into four subgroups according to the treatment they receive (the active drug or placebo) and how they respond to the treatment ('responders' or 'non-responders'): (1) Always Responders respond to both the active drug and placebo, (2) Placebo-only Responders respond to placebo but not to the active drug, (3) Drug-only Responders respond to the active drug but not to placebo, and (4) Never Responders do not respond to the active drug nor placebo. Subjects' response to the active drug and placebo are two latent characteristics and are predetermined as basic characteristics of the subject. Therefore, the enrolled population is the composite of these four types of subjects, with respective proportions of p_1 , p_2 , p_3 , and p_4 , as shown in Table 4.1.

Table 4.1: Distribution of the overall population

Placebo Group	Drug Group		
	Responders	Non-responders	
Responders	p_1	p_2	Placebo responders
Non-responders	p_3	p_4	Placebo non-responders
		Drug responders	Drug non-responders

As noted in the research of Muthén *et al.* [22], the prevalence of the four types of subjects in Table 4.1 is of clinical interest. Always Responders respond to both the active

drug and placebo. Both treatment therapies work for them. They may not need the active drug, but they may benefit more from the active drug than from placebo. Placebo-only Responders only respond to placebo, which may indicate side effects from the active drug for them. This is presumably a small group, or even with the proportion p_2 close to zero. Drug-only Responders only respond to the active drug, not to placebo. This group is of particular interest in the clinical trials, as this group may be considered as experiencing the real treatment effect. Chen *et al.* [5] also consider this group as the target population. They presume the ultimate goal of enrichment is to find this subgroup in the analysis population. Hopefully, the measure of treatment efficacy can be more readily reflected in this group of subjects. Both the investigational drug and placebo are not effective to Never Responders. They may be considered to switch to another drug.

In SED, the enrichment process is to first exclude ‘placebo responders’ in the placebo lead-in phase and then exclude ‘drug non-responders’ in the second stage. Theoretically, the enrichment process will identify the subjects who are Drug-only Responders, and we can hopefully get an unbiased estimate of treatment efficacy from this subgroup; however, in reality, it is not possible to find this pure target population in the sample. The subjects’ classification may differ from their real status due to random errors, measurement error, and subjective selection criteria [5]. Even though the subjects’ true status is used for data generation, their classification is solely based on the pre-determined criteria and the observed data.

We will still use notations from Chen’s manuscript [5] to describe the situation. It is natural to classify the subjects into two groups: ‘responders’ or ‘non-responders’, under the active drug or placebo, respectively. These two classifications should be independent, meaning that whether a subject responds to the active drug has nothing to do with whether this subject is a ‘placebo responder’. Therefore, we assume that the means of the change score from the beginning of the stage to the end of the stage is μ_R^d for ‘drug responders’ and μ_{NR}^d for ‘drug non-responders’ in the entire population under the active drug. Similarly,

we assume that the means of the change score from the beginning of the stage to the end of the stage is μ_R^p for ‘placebo responders’ and μ_{NR}^p for ‘placebo non-responders’ in the entire population under placebo. We assume the mean changes in these four subgroups keep the same from stage to stage. Then, we can describe the means of change score from the beginning of the stage to the end of the stage in the four subgroups as: $\mu_1 = \mu_R^d - \mu_R^p$, $\mu_2 = \mu_{NR}^d - \mu_R^p$, $\mu_3 = \mu_R^d - \mu_{NR}^p$, and $\mu_4 = \mu_{NR}^d - \mu_{NR}^p$. The specifications and derivations of the means of the change scores in the four subgroups are presented in Table 4.2.

Table 4.2: Means of the outcome changes from the baseline to the end of stage

Drug Responder	Drug Non-responder	Placebo Responder	μ_R^p
μ_R^d	μ_{NR}^d	Placebo Non-responder	μ_{NR}^p

Treatment Effect	Drug Responders	Drug Non-responders
Placebo Responders	$\mu_1 = \mu_R^d - \mu_R^p$	$\mu_2 = \mu_{NR}^d - \mu_R^p$
Placebo Non-responders	$\mu_3 = \mu_R^d - \mu_{NR}^p$	$\mu_4 = \mu_{NR}^d - \mu_{NR}^p$

Using the notations in Table 4.1 and Table 4.2, the overall treatment effect can then be defined as the sum of the products of the treatment effect and the corresponding proportion from the four sub-populations: $\mu = p_1 \times \mu_1 + p_2 \times \mu_2 + p_3 \times \mu_3 + p_4 \times \mu_4 = p_1 \times (\mu_R^d - \mu_R^p) + p_2 \times (\mu_{NR}^d - \mu_R^p) + p_3 \times (\mu_R^d - \mu_{NR}^p) + p_4 \times (\mu_{NR}^d - \mu_{NR}^p)$. The treatment effect in the target population is then $\mu_3 = \mu_R^d - \mu_{NR}^p$.

4.2.2 Concerns about the implementation of SED

In Chen’s numerical evaluation, they set the means of change scores for ‘drug responders’, ‘drug non-responders’, ‘placebo responders’, and ‘placebo non-responders’ as $\mu_R^d = -3.5$, $\mu_{NR}^d = -3.0$, $\mu_R^p = -3.0$, and $\mu_{NR}^p = -2.0$, respectively. The treatment effects for the four groups (Always Responders, Placebo-only Responders, Drug-only Responders, and

Never Responders) are -0.5, 0, -1.5, and -1, as indicated in Table 4.3. Under this setting, the treatment effect in the target population (Drug-only Responders) is -1.5. They have examined different settings on patient distribution by changing p_1 , p_2 , p_3 , and p_4 .

Table 4.3: Chen’s parameter selection

Treatment Effect	True Drug Responders	True Drug Non-responders
True placebo responders	$\mu_1 = \mu_R^d - \mu_R^p = -0.5$	$\mu_2 = \mu_{NR}^d - \mu_R^p = 0.0$
True placebo non-responders	$\mu_3 = \mu_R^d - \mu_{NR}^p = -1.5$	$\mu_4 = \mu_{NR}^d - \mu_{NR}^p = -1.0$

When they compared the results across the different designs: parallel design, placebo lead-in design, SPCD, TED, and SED, they assigned a fixed weight 0.7 to Stage I, and a weight 0.3 to Stage II of SPCD and SED, and an equal weight 0.15 to the second and third component of TED. To be specific, the test statistics for these five designs can be described as below. In parallel design and placebo lead-in design, $\hat{\delta}$ is the estimate of mean treatment difference in the randomized phase. In SPCD, $\hat{\delta}_1$ is the estimate of mean treatment difference in Stage I, and $\hat{\delta}_2$ is the estimate of mean treatment difference in ‘placebo non-responders’ in Stage II. In TED, $\hat{\delta}_1$ is the estimate of mean treatment difference in Stage I; $\hat{\delta}_2$ and $\hat{\delta}_3$ are the estimates of mean treatment difference in ‘placebo non-responders’ and ‘drug responders’ in Stage II. In SED, $\hat{\delta}_1$ is the estimate of mean treatment difference in ‘placebo non-responders’ in Stage I; and $\hat{\delta}_2$ is the estimate of mean treatment difference in ‘drug responders’ in Stage II. The weights ω_1 , ω_2 , and ω_3 sum up to one in each statistic.

$$Z_{\text{parallel}} = Z_{\text{lead-in}} = \frac{\hat{\delta}}{\sqrt{\text{Var}(\hat{\delta})}}$$

$$Z_{\text{SPCD}} = Z_{\text{SED}} = \frac{\omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_2}{\sqrt{\omega_1^2 \text{Var}(\hat{\delta}_1) + \omega_2^2 \text{Var}(\hat{\delta}_2)}}$$

$$Z_{\text{TED}} = \frac{\omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_2 + \omega_3 \hat{\delta}_3}{\sqrt{\omega_1^2 \text{Var}(\hat{\delta}_1) + \omega_2^2 \text{Var}(\hat{\delta}_2) + \omega_3^2 \text{Var}(\hat{\delta}_3)}}$$

Chen *et al.* compared the five designs by using a two-sample t -test with the simulated data. They concluded that SED shows the smallest bias in general but not the smallest MSE. However, when the proportion of the target population decreases, the gain of SED remains consistent compared with other designs. On the other side, they reported even though the power from SED ranks the second-highest after TED, the power differences between SED and TED are all within a minimal range (less or equal than 0.08).

SED shows some promising characteristics based on Chen’s numerical evaluation. However, the additional stage in SED is a cause for concern. Chen *et al.* claimed that the duration of the placebo lead-in stage is not necessarily the same as that of the succeeding stages. This may suggest that SED will not extend the whole process much longer than that of the two-stage designs. However, this claim conflicts with their conclusion in the meta-analysis about the placebo lead-in design they performed in 2011 [4]. In the meta-analysis, they reported a longer placebo lead-in phase related to a low placebo response rate. This implies extending the placebo lead-in phase can help the conduct of SED and reduce the placebo response rate. However, this additional long placebo lead-in phase will be a considerable burden on the pharmaceutical companies. Chen *et al.* didn’t provide a rationale for the parameter selection. After careful reviews, we believe that the evaluation should be based on more reasonable parameters. In addition, the previous research only reported the number of ‘placebo non-responders’ from Stage 0. It is not appropriate to consider this number as the required sample size, as it ignores the considerable proportion of ‘placebo responders’. Furthermore, it states the ultimate goal of SED is to increase the percentage of the target population through the enrichment process, but how much we will gain from it is uncertain. From all these perspectives, we believe that SED needs a further

evaluation from different angles before we can consider it as an option. In the following sections, we will critically appraise the performance of SED from three perspectives. We will first test the robustness of SED by varying the values of parameters. We will then calculate the actual sample size and the proportion of the target population in the sample. We will also explore the possibility of the application of new analysis methods in SED. The existing OLS method assumes independence between stages and implements hard cutoff to determine ‘placebo responders’ and ‘drug responders’. However, outcomes are collected multiple times in SED. It is natural to consider the outcomes from different stages as repeated measures for a subject. We will apply a repeated measures model [8] and a weighted repeated measures model [29] to SED, and compare their performance with the OLS method.

4.2.3 Data generation

As described in Section 4.2.1, the overall population is considered as a mixture of four subgroups. We assume the responses come from a mixture of a normal distribution with four components. The weights of the components equal to the proportion of each subgroup. The mixture distribution can be specified as $f(x) = \sum_{i=1}^4 p_i \phi_i$, where p_i is the proportion of each subset and ϕ_i is a density function from a normal distribution. The subjects’ actual status is generated from a multinomial distribution with probabilities p_1, p_2, p_3 , and p_4 for Always Responders, Placebo-only Responders, Drug-only Responders, and Never Responders, respectively. The sum of these probabilities is one, as they are corresponding to the proportions of the four subgroups.

We use the same assumptions as Chen *et al* [5]. The baseline response is assumed to be normally distributed with mean 25 and standard deviation one. The response at the end of a specific stage is generated based on the subject’s true status and the received treatment. A summary of parameters for the mean responses at the end of a stage is presented in Table 4.4. It is assumed the random error to be from normally distributed with mean zero and standard deviation four for different stages.

Table 4.4: Parameters for the mean responses

Subject True Status	Subject Received Treatment	
	Placebo	Drug
Always Responder	$\mu_0 + \mu_R^p$	$\mu_0 + \mu_R^d$
Placebo-only Responder	$\mu_0 + \mu_R^p$	$\mu_0 + \mu_{NR}^d$
Drug-only Responder	$\mu_0 + \mu_{NR}^p$	$\mu_0 + \mu_R^d$
Never Responder	$\mu_0 + \mu_{NR}^p$	$\mu_0 + \mu_{NR}^d$

We also have the following assumptions for data generation. The treatment effects within each subgroup are assumed the same from stage to stage. We assume that the generated data have the same variance in both stages. This restriction can be easily removed to incorporate different covariance structures. Re-randomization is implemented at the beginning of the specific stage; thus, the correlation between the estimates of treatment effect in different stages can be ignored. We will relax this restriction when generating the data for model construction in Section 4.3.4. The same criterion of a 10% reduction from the baseline, as Chen *et al.* selected, is used as the threshold for identification of ‘non-responders’ or ‘responders’. That is, if the reduction from baseline is over 10%, then the subject is considered as a ‘responder’, no matter whether the subject receives the active drug or placebo. As noted in Chen’s analysis [5], although the subjects’ true status is used to generate the responses, only the observed data are used to classify subjects as either ‘responders’ or ‘non-responders’. We generated 10,000 datasets to evaluate the different performance metrics and considered the following sample size: 50, 100, 150, and 200 subjects with 1:1 randomization to the active drug and placebo at the end of Stage 0.

4.3 Simulation Study

4.3.1 Parameter selection

In Chen’s setting, the mean baseline response is set to 25 ($\mu_0 = 25$). The threshold for the identification of ‘non-responders’ and ‘responders’ is a 10% reduction from the baseline value. This indicates the hard cutoff will be -2.5 if the baseline response is 25. Therefore, it comes to a simple conclusion that if the mean reduction of a subpopulation is larger than 2.5 ($\mu \leq -2.5$), subjects from this subpopulation will have a high probability of being a ‘responder’. If the mean reduction of a sub-population is smaller than 2.5 or increase happens in the mean change ($\mu > -2.5$), subjects from this sub-population will have a high probability of being a ‘non-responder’. As we suppose the change from baseline can be described by a normal distribution, in order to be classified as ‘placebo responders’, the mean change from baseline for ‘placebo responders’ should satisfy $\mu_R^p \leq -2.5$. On the other hand, if the mean change of a sub-population satisfies $\mu > -2.5$, subjects from this sub-population will have a high probability of being a ‘non-responder’. Similarly, in order to be classified as ‘drug non-responders’, the mean change from baseline for ‘drug non-responders’ should satisfy $\mu_{NR}^d \geq -2.5$. From these two constraints, we might have a situation that is $\mu_R^p \leq \mu_{NR}^d$.

In this subsection, we will reconsider parameter selection under three scenarios and see how parameter selection will affect the performance metric:

1. The original mean change in the target population is $\mu_3 = -1.5 > -2.5$. Based on the previous discussion, this setting will have more subjects as ‘non-responders’. This is not consistent with the expectation that this subgroup is Drug-only Responders. Therefore, we will evaluate situation $\mu_3 < -2.5$ in Setting 1, Setting 2, and Setting3;
2. In Chen’s setting, both μ_{NR}^d and μ_R^p were set to be equal and smaller than the average threshold. We will evaluate $\mu_{NR}^d = \mu_R^p = -2.0 > -2.5$ in Setting 2;
3. Chen’s setting sets μ_{NR}^d equals to μ_R^p . We will evaluate the case that μ_{NR}^d is different from μ_R^p in Setting 3.

To be specific, the three parameter settings are:

Alternative Setting 1: $\mu_R^d = -4.0$, $\mu_{NR}^d = -3.0$, $\mu_R^p = -3.0$, $\mu_{NR}^p = -1.0$;

Alternative Setting 2: $\mu_R^d = -4.0$, $\mu_{NR}^d = -2.0$, $\mu_R^p = -2.0$, $\mu_{NR}^p = -1.0$;

Alternative Setting 3: $\mu_R^d = -4.0$, $\mu_{NR}^d = -3.0$, $\mu_R^p = -2.0$, $\mu_{NR}^p = -1.0$.

We compared the five designs (parallel design, placebo lead-in, SPCD, TED, and SED) under these three selected parameter settings. From the simulation results, we noticed that TED always has higher power than SPCD and SED, especially when there is a large proportion of Drug-only Responders. SPCD and SED have similar levels of power with small sample sizes. The advantage of SED in power is noticeable only when we have a large sample size. It is not surprising that TED has superb performance in power. It uses more information from Stage II. Placebo ‘non-responders’ and ‘drug responders’ contribute twice in the estimate of treatment effect. When it comes to MSE, SED achieves smaller MSE than SPCD across different parameter selections and sample sizes, but not as small as TED. All these three designs have a similar bias. Therefore, it is not presented in the table. TED performs better than SED in all different scenarios. There is no reason to skip TED and select SED in the design phase considering the gain and loss. The detailed results are presented in Appendix D, Table D.2, Table D.3, and Table D.4. The results under the original setting reported by Chen *et al.* are presented in Table D.1.

4.3.2 Sample size determination

In Chen’s manuscript [5], they claimed when comparing different designs the sample size N refers to the number of subjects who were randomized at the first stage as opposed to subjects who were enrolled. For SED, these randomized subjects include only ‘placebo non-responders’ because ‘placebo responders’ will not enter the second stage nor be randomized. However, we believe that it is not appropriate to ignore those ‘placebo responders’ in the first stage, even though their data will not be used for the final analysis. The number of

subjects required for a study should include all the subjects who are enrolled and take at least one dose of the study medication, no matter it is the active drug or placebo. By using simulated data, we calculated the number of subjects needed for enrollment, which will ensure the target number of subjects for randomization. The percentage of subjects in the analysis set is defined as the number of subjects used in analysis divided by the number of subjects enrolled. It can be viewed as an index indicating the utility of enrolled subjects in SED. Table E.1 presents the results of the actual sample size and percentage of subjects in the analysis set from Chen’s parameter setting. Table E.2 presents the results from parameter Setting 1. As the proportion of subjects entering the second stage only depends on μ_{NR}^p and μ_R^p , Alternative Setting 2 and Alternative Setting 3 share the same results. We only present the results from Setting 2 in Table E.3.

In Chen’s setting in Table E.1, $\mu_R^p = -3.0 < -2.5$ but $\mu_{NR}^p = -2.0 > -2.5$. This leads to the lowest proportion of ‘placebo non-responders’ at the first stage in the four different parameter settings. Only about half of the enrolled subjects will enter the second stage and be used for the estimation of the treatment effect. In Alternative Setting 1, μ_R^p keeps the same as in Chen’s setting, but μ_{NR}^p is further away from the threshold than the original setting. Therefore, fewer subjects are needed at enrollment to achieve the same number of subjects entering the second stage. In Alternative Setting 2 and Alternative Setting 3, both μ_R^p and μ_{NR}^p are larger than the threshold -2.5, which leads to the highest proportion of ‘placebo non-responders’ in the first stage and the fewest number of subjects in need at enrollment. Therefore, the total sample size required is related to the placebo response rate. It is also noteworthy that, comparing to SPCD or TED, SED will need much more subjects enrolled at baseline, to achieve the same number of subjects for randomization. Considering the cost and time of subject enrollment, this will be a big limitation of SED.

4.3.3 Proportion of the target population in the sample

SED intends to get an unbiased estimate of the overall treatment effect by increasing the proportion of the target population in the sample from stage to stage. We estimate the proportion of the target population in the sample at the end of the study with different parameter settings under different population distributions. The results are presented in Table 4.5. Interestingly, the percentage change of the proportion of the target population in the sample is increasing with the decrease in the proportion of this subgroup. When the treated population includes a large proportion of subjects from the target population, there is no big change in the proportion of the target population in the sample over the enrichment process of SED. For example, if 90% of the subjects in the treated population comes from the target population, say Drug-only Responders, then the maximum percentage increase in the proportion of the target population is around 2% at the end of the study, as seen in Alternative Setting 2. When the proportion of the target population reduces to 40%, the percentage change between the true value and the estimated value largely increases, but is still below 25%, in the best scenario. In addition, the percentage change depends on parameter selection (response rate) and the original proportions of the four sub-populations. This indicates the enrichment process in SED is not robust.

4.3.4 New analysis methods

Outcomes are collected multiple times in SED. It is natural to consider the outcomes from different stages as repeated measures for a subject. Correlation between outcomes from different stages should also be taken into account in the estimation of drug efficacy. In this section, we will explore the possibility of describing the SED framework with models.

Repeated Measures Model

The start point of SED is the selected ‘placebo non-responders’ at the end of Stage 0. ‘Drug responders’ among the ‘placebo non-responders’ will be identified at the end of Stage

Table 4.5: Proportion estimates of the target population in the sample

p_1	p_2	p_3	p_4	Chen's Setting		Alternative Setting 1		Alternative Setting 2		Alternative Setting 3	
				Est.	% Change	Est.	% Change	Est.	% Change	Est.	% Change
0.0	0.0	0.9	0.1	0.906	0.7%	0.910	1.1%	0.920	2.2%	0.910	1.1%
0.1	0.0	0.8	0.1	0.822	2.8%	0.837	4.6%	0.833	4.1%	0.823	2.9%
0.1	0.1	0.7	0.1	0.739	5.6%	0.767	9.6%	0.759	8.4%	0.741	5.9%
0.1	0.1	0.6	0.2	0.638	6.3%	0.665	10.8%	0.666	11.0%	0.641	7.2%
0.2	0.1	0.5	0.2	0.543	8.6%	0.576	15.2%	0.567	13.4%	0.545	9.0%
0.2	0.1	0.4	0.3	0.437	9.3%	0.467	16.8%	0.464	16.0%	0.442	10.5%
0.3	0.1	0.4	0.2	0.444	11.0%	0.479	19.8%	0.462	15.5%	0.445	11.2%
0.3	0.2	0.4	0.1	0.454	13.5%	0.498	24.5%	0.470	17.5%	0.453	13.2%

Chen's setting: $\mu_R^d = -3.5$, $\mu_{NR}^d = -3.0$, $\mu_R^p = -3.0$, $\mu_{NR}^p = -2.0$;
Alternative Setting 1: $\mu_R^d = -4.0$, $\mu_{NR}^d = -3.0$, $\mu_R^p = -3.0$, $\mu_{NR}^p = -1.0$;
Alternative Setting 2: $\mu_R^d = -4.0$, $\mu_{NR}^d = -2.0$, $\mu_R^p = -2.0$, $\mu_{NR}^p = -1.0$;
Alternative Setting 3: $\mu_R^d = -4.0$, $\mu_{NR}^d = -3.0$, $\mu_R^p = -2.0$, $\mu_{NR}^p = -1.0$.

I. Therefore, we first consider the drug response as a measurable binary characteristic (“present” or “absent” status) and model the Stage I and Stage II outcomes by allowing the existence of correlation. In the proposed repeated measures model [8], we will use all available data collected on ‘placebo non-responders’ at the end of Stage 0, data at the end of Stage I, and data at the end of Stage II. However, the treatment effect will be estimated with outcomes on ‘placebo non-responders’ at the end of Stage 0, outcomes at the end of Stage I, and outcomes from ‘drug responders’ at the end of Stage II. In the following, we will list all the models that are used in this analysis approach. Linear regression will be used to model the outcomes and the correlation between stages.

At Stage I:

$$\text{Equation (1): } \Delta Y_{1i} = \alpha_{01} + \alpha_{11}Y_{1i} + \delta_1 G_{1i} + \epsilon_{1i}; \quad i = 1 : N_{PNR}$$

where ΔY_{1i} is the change in outcome from the end of Stage 0 to the end of Stage I on the selected ‘placebo non-responders’. Y_{1i} is the outcome at the end of Stage 0. G_{1i} is the treatment indicator in Stage I.

At Stage II:

$$\text{Equation (2): } \Delta Y_{2i} = \alpha_{02} + \alpha_{12}Y_{2i} + \delta_2 G_{2i} + \epsilon_{2i}; \quad i = 1 : n_{DR}$$

$$\text{Equation (3): } \Delta Y_{2i} = \alpha_{03} + \alpha_{13}Y_{2i} + \epsilon_{3i}; \quad i = (n_{DR} + 1) : (n_{DR} + n_{DNR})$$

$$\text{Equation (4): } \Delta Y_{2i} = \alpha_{04} + \alpha_{14}Y_{2i} + \epsilon_{4i}; \quad i = (n_{DNR} + n_{DR} + 1) : N_{PNR}$$

Equation (2) relates the change in outcome from the end of Stage I to the end of Stage II (ΔY_{2i}) to the outcome at the end of Stage I (Y_{2i}) and the new treatment assignment (G_{2i}) for ‘drug responders’. Equation (3) and Equation (4) present the relationship between the change in outcome from the end of Stage I to the end of Stage II and the outcome at the end of Stage I for ‘drug non-responders’ and Stage I placebo subjects.

It is assumed that the error terms $\{\epsilon_{1i}\}$, $\{\epsilon_{2i}\}$, $\{\epsilon_{3i}\}$, and $\{\epsilon_{4i}\}$ are independently and identically distributed across individuals. As subjects have outcomes recorded for both of the stages, the outcomes in the two stages are correlated within each subject. We assume the correlation is the same across the subjects, regardless of whether their data in Stage II are used for the final estimate of the treatment effect. Therefore, we have the following covariance matrices:

$$(\epsilon_{1i}, \epsilon_{ji}) \sim N \left[(0, 0), \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right]; \quad i = 1 : N_{PNR}, \quad j = 2 : 4$$

Suppose the treatment effect in ‘placebo non-responders’ in Stage I is δ_1 and the treatment effect in ‘drug responders’ in Stage II is δ_2 , then the contrast of interest is

$$\delta_\omega = \omega_1 \delta_1 + \omega_2 \delta_2; \quad \omega_1 + \omega_2 = 1$$

which represents the weighted average treatment effect in ‘placebo non-responders’ in Stage I and in ‘drug responders’ in Stage II. This is estimated by $\hat{\delta}_\omega = \omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_2$, with $\hat{\delta}_1$ and $\hat{\delta}_2$, the model-based estimates of δ_1 and δ_2 , respectively. A test for $H_0 : \delta_\omega = 0$ is based on the test statistic

$$T = \frac{\omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_2}{\sqrt{\omega_1^2 \text{Var}(\hat{\delta}_1) + \omega_2^2 \text{Var}(\hat{\delta}_2) + 2\omega_1 \omega_2 \text{Cov}(\hat{\delta}_1, \hat{\delta}_2)}}$$

where the variances and covariances can be derived from the models specified ahead. It is assumed that T will approximately follow the standard normal distribution under the null hypothesis H_0 .

Weighted Repeated Measures Model

In addition, we consider the drug response as a characteristic that exists in every subject to a certain degree in this study. It is treated as a continuous measure, ranging from 1 to 0, from absolute ‘responder’ to absolute ‘non-responder’. We will propose to include it in the model as a weight for the drug group. The full weighted repeated measure model [29] can be specified as below:

At Stage I:

$$\text{Equation (1): } \Delta Y_{1i} = \alpha_{01} + \alpha_{11} Y_{1i} + \delta_1 G_{1i} + \epsilon_{1i}; \quad i = 1 : N_{PNR}$$

At Stage II:

$$\text{Equation (2): } \Delta Y_{2i} = \alpha_{02} + \alpha_{12} Y_{2i} + \delta_2 G_{2i} + \epsilon_{2i}; \quad i = 1 : n_D$$

$$\text{Equation (3): } \Delta Y_{2i} = \alpha_{03} + \alpha_{13} Y_{2i} + \epsilon_{3i}; \quad i = (n_D + 1) : N_{PNR}$$

The first equation relates the change in outcome from baseline to the end of Stage I (ΔY_{1i}) to the outcome at the end of Stage 0 (Y_{1i}) and the treatment assignment during Stage I (G_{1i}) in the selected ‘placebo non-responders’. The second equation relates the change in outcome from the end of Stage I to the end of Stage II (ΔY_{2i}) to the outcome at the end of Stage I (Y_{2i}) and the new treatment assignment (G_{2i}) for the active drug group in Stage II. The last equation relates the change in outcome from the end of Stage I to the end of Stage II (ΔY_{2i}) to the outcome at the end of Stage I (Y_{2i}) for Stage I placebo subjects.

As the outcome is normally distributed, the errors ϵ in the model are also normally distributed with mean $E(\epsilon) = \mathbf{0}$ and variance $\text{Var}(\epsilon) = \sigma^2 \mathbf{\Sigma}$, where σ^2 is unknown, and $\mathbf{\Sigma}$ is defined as follows, reflecting the correlation between Stage I and Stage II (ρ_{12}).

$$\mathbf{\Sigma}_i = v_i^{-1/2} \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix} v_i^{-1/2}; \quad i = 1 : n_D$$

$$\mathbf{\Sigma}_i = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix} \quad i = n_D : N_{PNR}$$

The weights are set to 1 for all ‘placebo non-responders’ in Stage I. In Stage II, all subjects in the Stage I placebo group are assigned to a weight of 1, but subjects in the Stage I active drug group are assigned weights based on their response to the active drug.

Therefore, v_i takes the following form:

$$\mathbf{v}_i = \begin{bmatrix} 1 & 0 \\ 0 & v_i \end{bmatrix} \quad i = 1 : n_D$$

In matrix form, the model can be written as $\mathbf{\Delta Y}_i = \mathbf{X}_i \boldsymbol{\delta} + \epsilon_i$, where $\mathbf{\Delta Y}_i$ is a vector of outcomes and \mathbf{X}_i is the covariate matrix for individual i . The generalized least squares estimate for the vector of coefficients is $\hat{\boldsymbol{\delta}} = \{\sum_{i=1}^N \mathbf{X}'_i \mathbf{\Sigma}_i^{-1} \mathbf{X}_i\}^{-1} \sum_{i=1}^N \mathbf{X}'_i \mathbf{\Sigma}_i^{-1} \mathbf{\Delta Y}_i$. With $\mathbf{\Sigma}_i$ known, the variance of the estimate is $\text{Var}(\hat{\boldsymbol{\delta}}) = \sigma^2 \{\sum_{i=1}^N \mathbf{X}'_i \mathbf{\Sigma}_i^{-1} \mathbf{X}_i\}^{-1}$.

Therefore, the contrast of interest can be specified as:

$$\delta_\omega = \omega_1 \delta_1 + \omega_2 \delta_2; \quad \omega_1 + \omega_2 = 1$$

It is estimated by $\hat{\delta}_\omega = \omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_2$, with $\hat{\delta}_1$ and $\hat{\delta}_2$, the model-based estimates of δ_1 and δ_2 , respectively. A test for $H_0 : \delta_\omega = 0$ is based on the test statistic

$$T = \frac{\omega_1 \hat{\delta}_1 + \omega_2 \hat{\delta}_2}{\sqrt{\omega_1^2 \text{Var}(\hat{\delta}_1) + \omega_2^2 \text{Var}(\hat{\delta}_2) + 2\omega_1 \omega_2 \text{Cov}(\hat{\delta}_1, \hat{\delta}_2)}}$$

where the treatment effects, variances, and covariances are estimated from the model specified subsequently. It is assumed that T follows approximately the standard normal distri-

bution under the null hypothesis.

As discussed by Rybin *et al.* [29], the zero-to-one scale can be derived in different ways. The easiest way might be through a logistic regression. A logistic regression is constructed with baseline characteristics, and the outcomes at both baseline and at the end of Stage I. The probability of being a ‘drug responder’ is estimated through the model:

$$\text{logit}(p_{R_i=1}) = \alpha + \beta_0 Y_{1i} + \beta_1 X_i; \quad i = 1 : n_D$$

where Y_{1i} is the outcome at the end of Stage 0 for subject i , X_i is another baseline characteristic, R_i is a response indicator with a value of 1 or 0 (for ‘drug responder’ or ‘drug non-responder’, respectively), and n_D is the number of subjects in the active drug group at Stage I.

Comparisons of different approaches

We assessed the performance of the OLS method, the repeated measures model, and the weighted repeated measures model with the simulated data in the SED framework. We considered the following sample sizes: 50, 100, 150, and 200 subjects with 1:1 randomization at the end of Stage 0 to the active drug and placebo. Correlations between the change in outcome during Stage I and the change in the outcome during Stage II were assumed to be the same for all treatment arms and equal to -0.5, -0.4, -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4, and 0.5. The values of $\mu_{R^d}^d$, $\mu_{NR^d}^d$, $\mu_{R^p}^p$, and $\mu_{NR^p}^p$ are set to -3.5, -3.0, -3.0, and -2.0 under the alternative hypothesis. The target treatment effect μ_3 is -1.5 in this setting. Under the null hypothesis, the target treatment effect is set to 0. All the other parameters are set to the same values as in Chen’s manuscript. We generated 10,000 datasets to evaluate the different performance metrics. As shown in Figure 4.2, the weighted repeated measures model has the lowest type I error for all the sample sizes. It is well controlled below or around 0.05. This is also correct across all the correlation values. The repeated measures model and the OLS method have a similar performance in type I error. Figure 4.3 presents the MSE of the

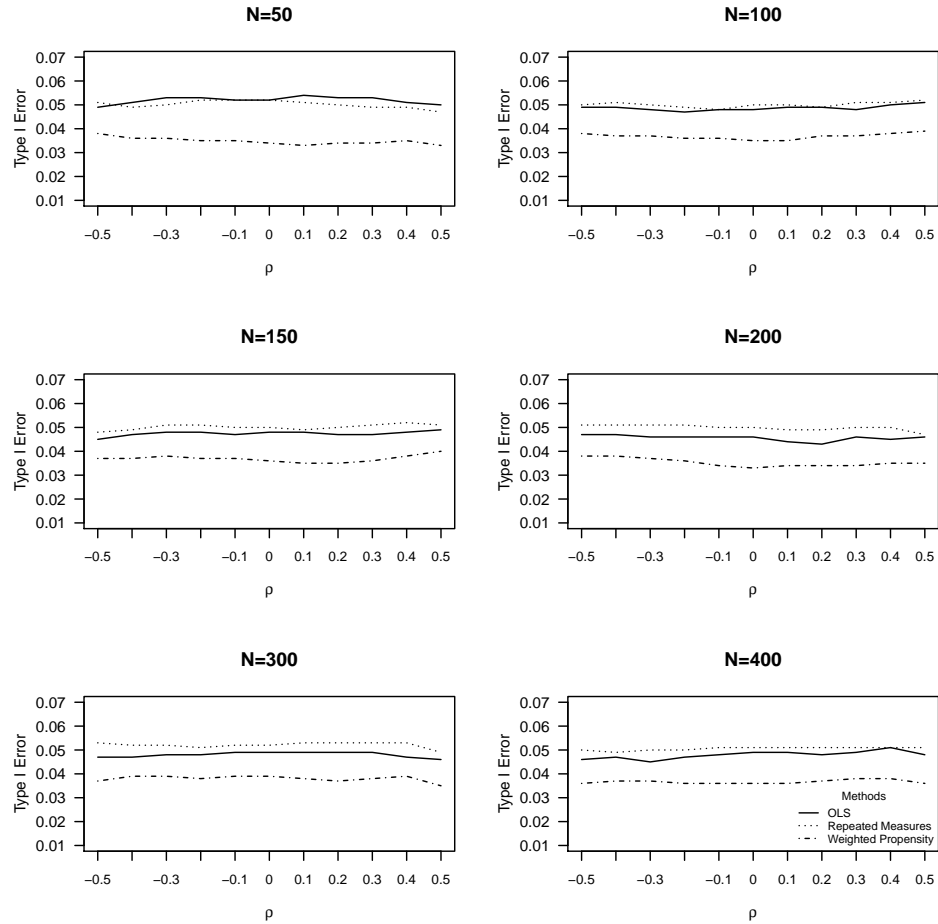


Figure 4.2: Type I error within the framework of SED

treatment effect. When the sample size is small, OLS has a smaller MSE compared to the two repeated measures models, and the estimate is consistent across different correlation values. The difference in MSE among the three methods gets smaller and smaller, with the increase of the sample size. While the weighted repeated measures model has lower power, the difference is trivial (Figure 4.4).

4.4 Conclusions

The SED was first proposed by Chen *et al.* in 2014, as an attempt to improve the estimation of the target treatment effect for clinical trials that have a mixed population, assuming the existence of ‘placebo responders’ [5]. It aims to exclude not only ‘placebo responders’, but

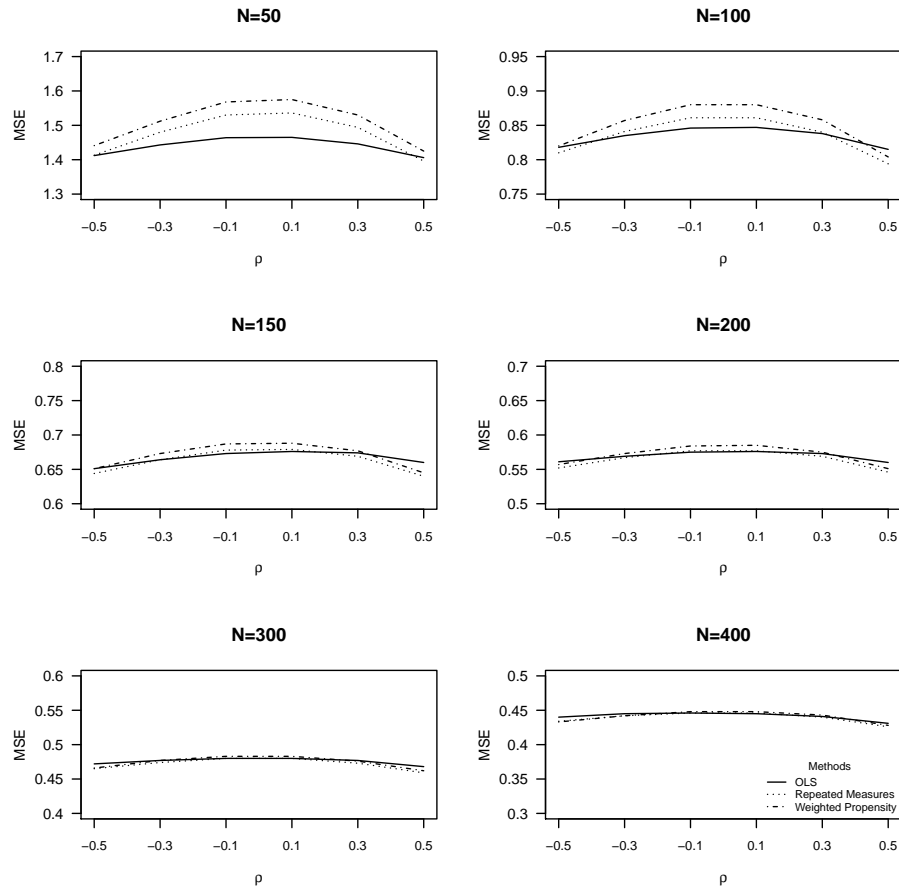


Figure 4-3: MSE within the framework of SED

also ‘drug non-responders’, in a sequential manner. If there is a considerable proportion of either ‘placebo responders’ or who are categorically Never Responders in the population, the existing enrichment designs in clinical trials, like SPCD and TED, may introduce bias to the estimate of treatment effect as they don’t exclude the ‘drug non-responders’. In their simulation results, it has been demonstrated that SED has a smaller bias in the estimation of treatment effect and yields reasonably high power in general.

However, in this chapter, we appraised SED from different aspects and conclude that the additional stage might not be critical. First of all, we discussed the parameter selection. In the original manuscript, the authors selected a 10% reduction from the baseline response as the threshold for identifying ‘responders’ and only used one set of parameters. In our

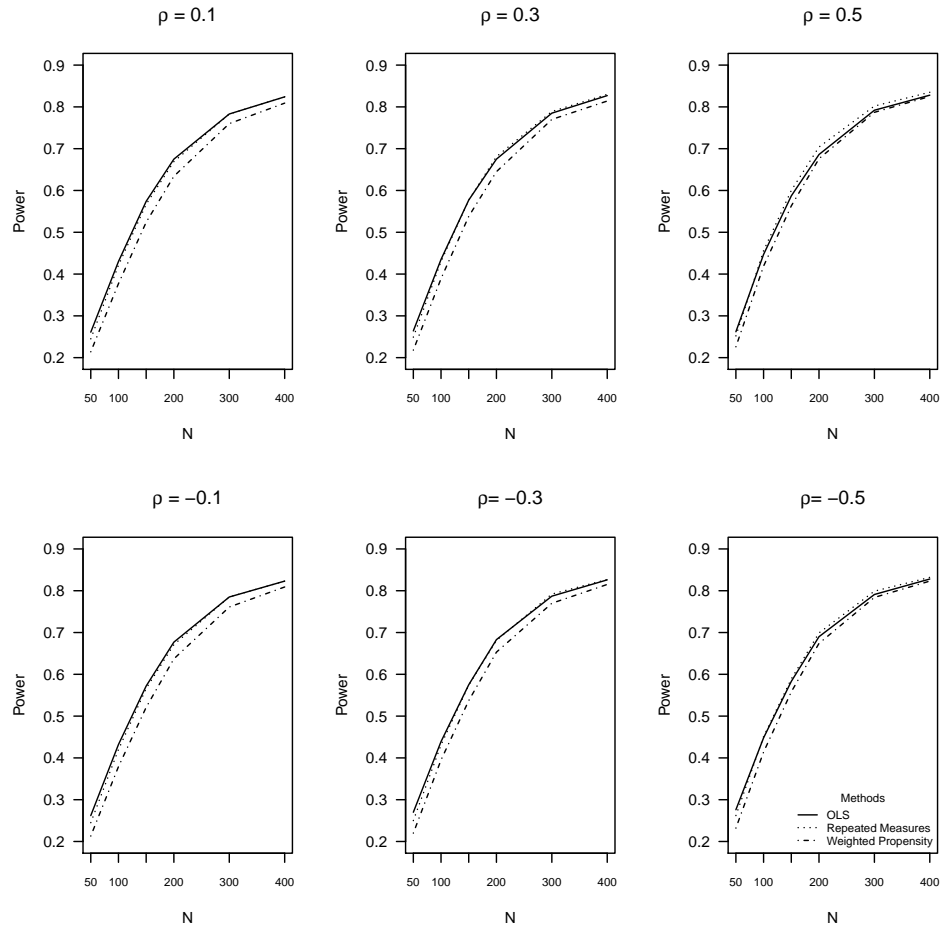


Figure 4-4: Power within the framework of SED

evaluation, we provided three different parameter settings, to test the robustness of SED to various proportions of ‘responders’ and ‘non-responders’. Interestingly, when the sample size is large, there is no significant difference in power across the five different designs. But in reality, we might not want to run such a large trial in psychiatry studies. TED shows advantages over SED in power in small sample size, as it uses more information from the second stage. ‘Placebo non-responders’ and ‘drug responders’ contribute twice in the estimate of treatment effect. SED achieves smaller MSE than SPCD, but not as small as TED. When it comes to bias, none of these designs is superior to the others. All these performance metrics indicate SED is inferior to TED. There is no reason to skip TED and implement SED directly.

Then, we evaluated the actual number of subjects needed at the enrollment. This is a substantial drawback of SED that many more subjects need to be recruited and screened to ensure enough ‘placebo non-responders’ entering the randomization stage. As it was noted by Chen *et al.*, the number of subjects recruited will be approximately the number of subjects randomized times the reciprocal of the ‘placebo non-response rate [5]. The placebo non-response rate is difficult to estimate in practice; therefore we used the simulated data to calculate the actual number of subjects needed at the enrollment. We evaluated different parameter settings and various distributions of the population. The proportion of subjects entering the randomization phase is no more than two-thirds of the whole population who are enrolled in the placebo lead-in phase. It will be a non-ignorable issue that SED requires a higher cost in subject enrollment. Due to the nature of this design, all enrolled subjects take placebo at the first stage, which may lead to a high dropout rate. In return, it will require even more subjects recruited at the beginning of the study. Furthermore, this design has one more stage than SPCD and TED. The additional stage will largely extend the duration of clinical trials, which will be a burden to the pharmaceutical company and the investigational sites. Based on the report done by Nutt and Goodwin [23], it is very time-consuming (13 years in average in new drug development) and expensive (estimates range from \$800 million to \$3 billion per new agent) for pharmaceutical companies to introduce a new CNS drug to the market.

Next, we estimated the proportion of the target population in the sample at the end of the study. The SED intends to increase the percentage of the target population through a sequential enrichment scheme to get an unbiased estimate of the treatment effect. However, the simulation results do not provide reliable evidence to support this hypothesis. The premise of SED is the existence of a considerable proportion of Drug-only Responders in the population. However, the enrichment rate is very low for high portions of the target population in the treated subjects. It might be reasonable that the marginal increment decreases with the increase of the ratio of the target population. However, that will then

lose the meaning of the application of SED. Another issue with this enrichment process is that it largely depends on parameter selection and the distribution of the four sub-groups in the whole population. The simulation study only covers limited possibilities of parameter settings and distributions of subgroups, but the enrichment rates still have a wide range, from 0.7% (Chen's setting) to 24.5% (Alternative Setting 1). This phenomenon indicates the idea of enrichment in SED is not robust.

Lastly, we applied new analysis methods, including a repeated measures model and a weighted repeated measures model, to the SED framework. The weighted repeated measures model has an advantage in type I error over the other two methods. The OLS method provides an alternative option, as the type I error is well controlled below or around 0.05 in almost scenarios. The investigators might want to choose the OLS method as the implementation and calculation are relatively straightforward, compared to the weighted repeated measures model. In terms of MSE of the treatment effect and power, none of the three methods stands out. This also provides flexibility to the investigators, depending on how much prior information they have about the data.

As the only three-stage design, SED doesn't gain substantial benefits from the additional stage over the other two-stage designs, SPCD and TED. Instead, it requires a large sample size and extended trial duration. Even though it might have some charming performances, we still suggest investigators considering the two-stage enrichment designs when placebo response is a critical issue in practice.

Chapter 5

Summary and Future Studies

5.1 Summary

High placebo response can be a reason for the failure of clinical trials, especially in some fields related to the dysfunction of the central nervous system. The impact of placebo response to psychiatric disorders has been noted for over 50 years. However, the mechanism and the extent of the effect of placebo response are still not clear. The existence of placebo response makes the powerful tool, double-blinded randomization, not as powerful as before. Several multi-stage designs have been proposed in the past 20 years, trying to reduce the impact of placebo response and increase the possibility of successful trials.

We put our focus on two multi-stage designs, TED and SED, in this dissertation. These two designs are more complicated than the pioneer design SPCD. They have yet been implemented in practice. The concerns might come from the lack of flexible analysis methods and the complexity of the designs. We explored the possibility of application of the new analysis methods and the relationship between complexity and efficacy in these designs. TED is a two-stage design, with the same length as SPCD, but it also includes ‘drug responders’ in consideration of drug efficacy. SED, different from SPCD and TED, is a three-stage design. One placebo lead-in stage is added prior to the randomization phase.

Under the framework of TED, we first considered the placebo response as a measurable binary characteristic (‘present’ or ‘absent’ status) and modeled the Stage I and Stage II outcomes by allowing the existence of correlation. We proposed to use a repeated measures model to estimate the treatment efficacy. Then, we relaxed the first assumption and considered the placebo response as a characteristic that exists in every subject to a certain degree in this study. It is treated as a continuous measure, ranging from 0 to 1, from absolute ‘responder’ to absolute ‘non-responder’. We proposed to include it as a weight in the weighted repeated measures model for the placebo group. Lastly, we considered placebo response (also drug response in the active drug group) as a binary ‘present’ or ‘absent’ characteristic but now introduced a stochastic component in the classification of the subjects. We proposed to use the EM algorithm to estimate the parameters and latent characteristics. Based

on the simulation study, all these three approaches have some preferable characteristics. If the investigators have prior information about the relationship between placebo response and subject characteristics, the weighted repeated measures model will be a good choice, as it uses more measurements to evaluate a subject's probability of being a 'placebo non-responder'. If there is no prior information, the investigators can use the repeated measures model as an alternative. The EM algorithm will be extremely useful given an ambiguous definition of clinical response to placebo or the active drug. It also allows the investigators to have different response probabilities for placebo and the active drug. However, the complexity and limitations also increase progressively from the repeated measures model to the weighted repeated measures model to the EM algorithm. Without additional information, we can only derive a pseudo weight, which might impact the performance of the weighted repeated measures model. The EM algorithm takes a long time to get a converge result if it exists. It also requires a large sample size, which might be a big challenge for pharmaceutical companies in real clinical trials. Additionally, the EM algorithm might not be a wise choice if the means are similar in 'responders' and 'non-responders'. Therefore, the investigators should comprehensively evaluate and assess the prior information during the design phase to determine the appropriate analysis method.

SED, as the only three-stage design to date, is claimed to have some promising characteristics. However, after our further evaluation from parameter selection, sample size determination, and the increment of the proportion of the target population in the analysis sample, we believe SED might not be a good choice in the trial design. The additional stage doesn't tremendously improve the performance metrics. Instead, it largely extends the duration of the clinical trials and requires many more subjects at enrollment than its two-stage competitors. We suggest investigators focusing on two-stage designs.

Increasing the complexity of design can always lead a 'better' performance in some perspectives. The estimate of treatment efficacy can always be less biased on a more purified subset in the analysis population. But the challenge is whether we want to take the cor-

responding cost for the marginal improvement. We believe our research can provide some light on the exploration of better clinical design and flexible analysis methods when high placebo response is a critical issue in the clinical trials.

5.2 Future Studies

This dissertation work can be continued in several specific directions. We list these possible extensions here in the order the methods were presented above.

In the first proposed repeated measures models, more information can be collected in the determination of subjects' status. Multivariate logistic regression or multiple dimension clustering can be the possible tools towards this goal.

In the proposed weighted repeated measures model, a similar weight can be assigned to the active drug group. Subjects will be more like a 'drug responder' when the weight is closer to one, while a weight close to zero indicates a higher probability of being a 'drug non-responder'. Different covariates can be included in the logistic regression in determining the weights for the placebo group and the active drug group.

The latent dichotomous status approach can be further explored by considering different distributions of placebo response and drug response. For example, Beta distribution is for placebo response and Binomial distribution for drug response; or placebo response and drug response come from the same distribution family, but with different parameters. The proposed method works well in case of the good separation of the outcome distribution in 'responders' and 'non-responders'. Future research is needed to quantify the method efficiency for different degrees of separation.

Appendix A

M-step in EM Algorithm

All model parameters can be defined explicitly by differentiating Q-function.

Placebo Non-responders Stage I:

$$b_{01} = \frac{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})(y_{1i} - b_{11}y_{01i})}{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})}$$

$$b_{11} = \frac{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})y_{01i}(y_{1i} - b_{01})}{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})y_{01i}^2}$$

$$\sigma_{101}^2 = \frac{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})(y_{1i} - b_{01} - b_{11}y_{01i})^2}{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})}$$

Placebo Non-responders Stage II:

$$b_{02} = \frac{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})(y_{2i} - b_{12}y_{1i} - b_{22}g_{2i})}{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})}$$

$$b_{12} = \frac{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})y_{1i}(y_{2i} - b_{02} - b_{22}g_{2i})}{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})y_{1i}^2}$$

$$b_{22} = \frac{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})g_{2i}(y_{2i} - b_{02} - b_{12}y_{1i})}{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})g_{2i}^2}$$

$$\sigma_{201}^2 = \frac{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})(y_{2i} - b_{02} - b_{12}y_{1i} - b_{22}g_{2i})^2}{\sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})}$$

Placebo Responders Stage I:

$$b_{03} = \frac{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i} (y_{1i} - b_{13} y_{01i})}{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i}}$$

$$b_{13} = \frac{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i} y_{01i} (y_{1i} - b_{03})}{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i} y_{01i}^2}$$

$$\sigma_{102}^2 = \frac{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i} (y_{1i} - b_{03} - b_{13} y_{01i})^2}{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i}}$$

Placebo Responders Stage II:

$$b_{04} = \frac{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i} (y_{2i} - b_{14} y_{1i} - b_{24} g_{2i})}{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i}}$$

$$b_{14} = \frac{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i} y_{1i} (y_{2i} - b_{04} - b_{24} g_{2i})}{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i} y_{1i}^2}$$

$$b_{24} = \frac{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i} g_{2i} (y_{2i} - b_{04} - b_{14} y_{1i})}{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i} g_{2i}^2}$$

$$\sigma_{202}^2 = \frac{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i} (y_{2i} - b_{04} - b_{14} y_{1i} - b_{24} g_{2i})^2}{\sum_{i=1}^N (1 - g_{1i}) \pi_{1i}}$$

Drug Non-responders Stage I:

$$b_{05} = \frac{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})(y_{1i} - b_{15}y_{01i})}{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})}$$

$$b_{15} = \frac{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})y_{01i}(y_{1i} - b_{05})}{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})y_{01i}^2}$$

$$\sigma_{103}^2 = \frac{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})(y_{1i} - b_{05} - b_{15}y_{01i})^2}{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})}$$

Drug Non-responders Stage II:

$$b_{06} = \frac{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})(y_{2i} - b_{16}y_{1i} - b_{26}g_{2i})}{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})}$$

$$b_{16} = \frac{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})y_{1i}(y_{2i} - b_{06} - b_{26}g_{2i})}{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})y_{1i}^2}$$

$$b_{26} = \frac{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})g_{2i}(y_{2i} - b_{06} - b_{16}y_{1i})}{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})g_{2i}^2}$$

$$\sigma_{203}^2 = \frac{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})(y_{2i} - b_{06} - b_{16}y_{1i} - b_{26}g_{2i})^2}{\sum_{i=1}^N g_{1i}(1 - \pi_{2i})}$$

Drug Responders Stage I:

$$b_{07} = \frac{\sum_{i=1}^N g_{1i}\pi_{2i}(y_{1i} - b_{17}y_{01i})}{\sum_{i=1}^N g_{1i}\pi_{2i}}$$

$$b_{17} = \frac{\sum_{i=1}^N g_{1i}\pi_{2i}y_{01i}(y_{1i} - b_{07})}{\sum_{i=1}^N g_{1i}\pi_{2i}y_{01i}^2}$$

$$\sigma_{104}^2 = \frac{\sum_{i=1}^N g_{1i}\pi_{2i}(y_{1i} - b_{07} - b_{17}y_{01i})^2}{\sum_{i=1}^N g_{1i}\pi_{2i}}$$

Drug Responders Stage II:

$$b_{08} = \frac{\sum_{i=1}^N g_{1i}\pi_{2i}(y_{2i} - b_{18}y_{1i} - b_{28}g_{2i})}{\sum_{i=1}^N g_{1i}\pi_{2i}}$$

$$b_{18} = \frac{\sum_{i=1}^N g_{1i}\pi_{2i}y_{1i}(y_{2i} - b_{08} - b_{28}g_{2i})}{\sum_{i=1}^N g_{1i}\pi_{2i}y_{1i}^2}$$

$$b_{28} = \frac{\sum_{i=1}^N g_{1i}\pi_{2i}g_{2i}(y_{2i} - b_{08} - b_{18}y_{1i})}{\sum_{i=1}^N g_{1i}\pi_{2i}g_{2i}^2}$$

$$\sigma_{204}^2 = \frac{\sum_{i=1}^N g_{1i}\pi_{2i}(y_{2i} - b_{08} - b_{18}y_{1i} - b_{28}g_{2i})^2}{\sum_{i=1}^N g_{1i}\pi_{2i}}$$

Appendix B

Formulas for Variance Components in EM Algorithm

The covariance matrix of the parameters can be estimated with second derivatives of Q-function.

Placebo Non-responders Stage I:

$$\frac{\partial^2 Q}{\partial b_{01}^2} = -\frac{1}{\sigma_{101}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i}); \quad \frac{\partial^2 Q}{\partial b_{11}^2} = -\frac{1}{\sigma_{101}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})y_{01i}^2;$$

$$\frac{\partial^2 Q}{\partial b_{01}\partial b_{11}} = -\frac{1}{\sigma_{101}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})y_{01i}$$

Placebo Non-responders Stage II:

$$\frac{\partial^2 Q}{\partial b_{02}^2} = -\frac{1}{\sigma_{201}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i}); \quad \frac{\partial^2 Q}{\partial b_{12}^2} = -\frac{1}{\sigma_{201}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})y_{1i}^2;$$

$$\frac{\partial^2 Q}{\partial b_{22}^2} = -\frac{1}{\sigma_{201}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})g_{2i}^2; \quad \frac{\partial^2 Q}{\partial b_{02}b_{12}} = -\frac{1}{\sigma_{201}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})y_{1i}$$

$$\frac{\partial^2 Q}{\partial b_{02}b_{22}} = -\frac{1}{\sigma_{201}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})g_{2i}; \quad \frac{\partial^2 Q}{\partial b_{12}b_{22}} = -\frac{1}{\sigma_{201}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{1i})y_{1i}g_{2i}$$

Placebo Responders Stage I:

$$\frac{\partial^2 Q}{\partial b_{03}^2} = -\frac{1}{\sigma_{102}^2} \sum_{i=1}^N (1 - g_{1i})\pi_{1i}; \quad \frac{\partial^2 Q}{\partial b_{13}^2} = -\frac{1}{\sigma_{102}^2} \sum_{i=1}^N (1 - g_{1i})\pi_{1i}y_{01i}^2;$$

$$\frac{\partial^2 Q}{\partial b_{03}\partial b_{13}} = -\frac{1}{\sigma_{102}^2} \sum_{i=1}^N (1 - g_{1i})\pi_{1i}y_{01i}$$

Placebo Responders Stage II:

$$\frac{\partial^2 Q}{\partial b_{04}^2} = -\frac{1}{\sigma_{202}^2} \sum_{i=1}^N (1 - g_{1i})\pi_{1i}; \quad \frac{\partial^2 Q}{\partial b_{14}^2} = -\frac{1}{\sigma_{202}^2} \sum_{i=1}^N (1 - g_{1i})\pi_{1i}y_{1i}^2;$$

$$\frac{\partial^2 Q}{\partial b_{24}^2} = -\frac{1}{\sigma_{202}^2} \sum_{i=1}^N (1 - g_{1i})\pi_{1i}g_{2i}^2; \quad \frac{\partial^2 Q}{\partial b_{04}b_{14}} = -\frac{1}{\sigma_{202}^2} \sum_{i=1}^N (1 - g_{1i})\pi_{1i}y_{1i};$$

$$\frac{\partial^2 Q}{\partial b_{04}b_{24}} = -\frac{1}{\sigma_{202}^2} \sum_{i=1}^N (1 - g_{1i})\pi_{1i}g_{2i}; \quad \frac{\partial^2 Q}{\partial b_{14}b_{24}} = -\frac{1}{\sigma_{202}^2} \sum_{i=1}^N (1 - g_{1i})\pi_{1i}y_{1i}g_{2i}$$

Drug Non-responders Stage I:

$$\frac{\partial^2 Q}{\partial b_{05}^2} = -\frac{1}{\sigma_{103}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{2i}); \quad \frac{\partial^2 Q}{\partial b_{15}^2} = -\frac{1}{\sigma_{103}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{2i})y_{01i}^2$$

$$\frac{\partial^2 Q}{\partial b_{05} \partial b_{15}} = -\frac{1}{\sigma_{103}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{2i})y_{01i}$$

Drug Non-responders Stage II:

$$\frac{\partial^2 Q}{\partial b_{06}^2} = -\frac{1}{\sigma_{203}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{2i}); \quad \frac{\partial^2 Q}{\partial b_{16}^2} = -\frac{1}{\sigma_{203}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{2i})y_{1i}^2$$

$$\frac{\partial^2 Q}{\partial b_{26}^2} = -\frac{1}{\sigma_{203}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{2i})g_{2i}^2; \quad \frac{\partial^2 Q}{\partial b_{06} b_{16}} = -\frac{1}{\sigma_{203}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{2i})y_{1i}$$

$$\frac{\partial^2 Q}{\partial b_{06} b_{26}} = -\frac{1}{\sigma_{203}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{2i})g_{2i}; \quad \frac{\partial^2 Q}{\partial b_{16} b_{26}} = -\frac{1}{\sigma_{203}^2} \sum_{i=1}^N (1 - g_{1i})(1 - \pi_{2i})y_{1i}g_{2i}$$

Drug Responders Stage I:

$$\frac{\partial^2 Q}{\partial b_{07}^2} = -\frac{1}{\sigma_{104}^2} \sum_{i=1}^N g_{1i}\pi_{2i}; \quad \frac{\partial^2 Q}{\partial b_{17}^2} = -\frac{1}{\sigma_{104}^2} \sum_{i=1}^N g_{1i}\pi_{2i}y_{01i}^2$$

$$\frac{\partial^2 Q}{\partial b_{07} \partial b_{17}} = -\frac{1}{\sigma_{104}^2} \sum_{i=1}^N g_{1i}\pi_{2i}y_{01i}$$

Drug Responders Stage II:

$$\frac{\partial^2 Q}{\partial b_{08}^2} = -\frac{1}{\sigma_{204}^2} \sum_{i=1}^N g_{1i}\pi_{2i}; \quad \frac{\partial^2 Q}{\partial b_{18}^2} = -\frac{1}{\sigma_{204}^2} \sum_{i=1}^N g_{1i}\pi_{2i}y_{1i}^2$$

$$\frac{\partial^2 Q}{\partial b_{28}^2} = -\frac{1}{\sigma_{204}^2} \sum_{i=1}^N g_{1i}\pi_{2i}g_{2i}^2; \quad \frac{\partial^2 Q}{\partial b_{08} b_{18}} = -\frac{1}{\sigma_{204}^2} \sum_{i=1}^N g_{1i}\pi_{2i}y_{1i}$$

$$\frac{\partial^2 Q}{\partial b_{08} b_{28}} = -\frac{1}{\sigma_{204}^2} \sum_{i=1}^N g_{1i}\pi_{2i}g_{2i}; \quad \frac{\partial^2 Q}{\partial b_{18} b_{28}} = -\frac{1}{\sigma_{204}^2} \sum_{i=1}^N g_{1i}\pi_{2i}y_{1i}g_{2i}$$

Appendix C

Robustness Analyses in EM Algorithm

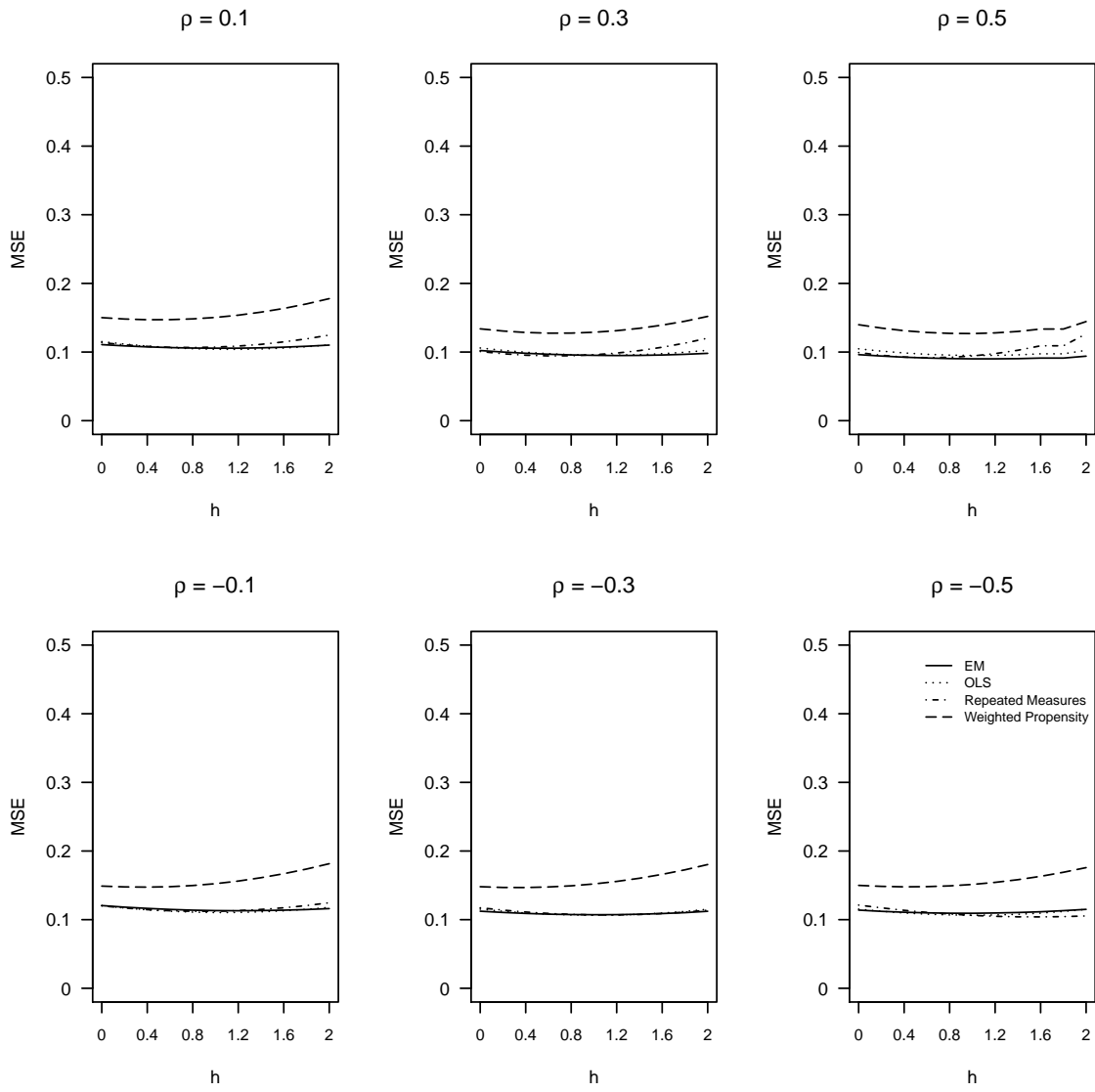


Figure C.1: MSE assuming correct response threshold specification - a fixed Stage II effect in ‘non-responders’ and various Stage II effects in ‘responders’

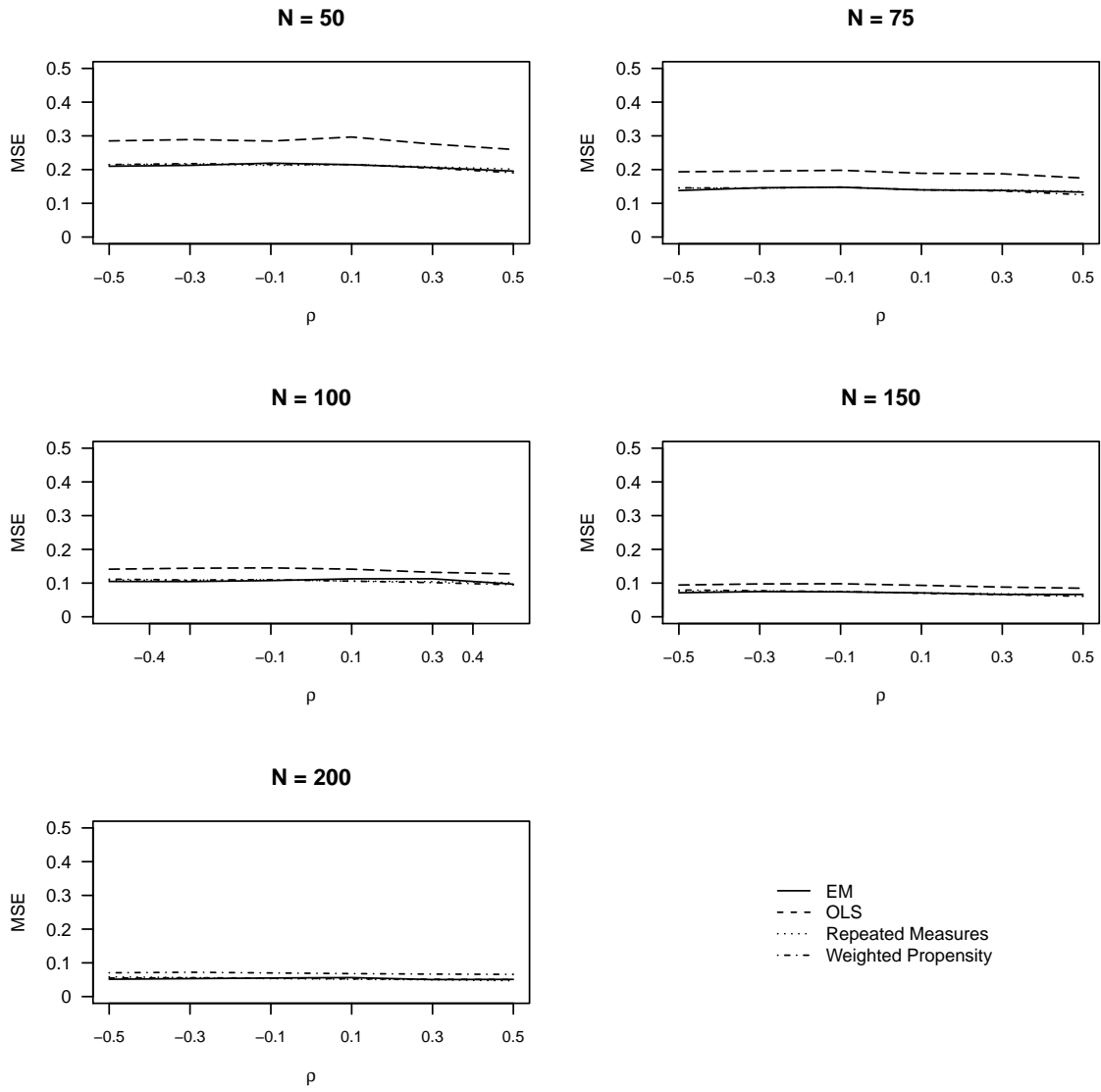


Figure C.2: MSE assuming correct response threshold specification - Stage II effect in ‘responders’ set to half of the Stage II effect in ‘non-responders’

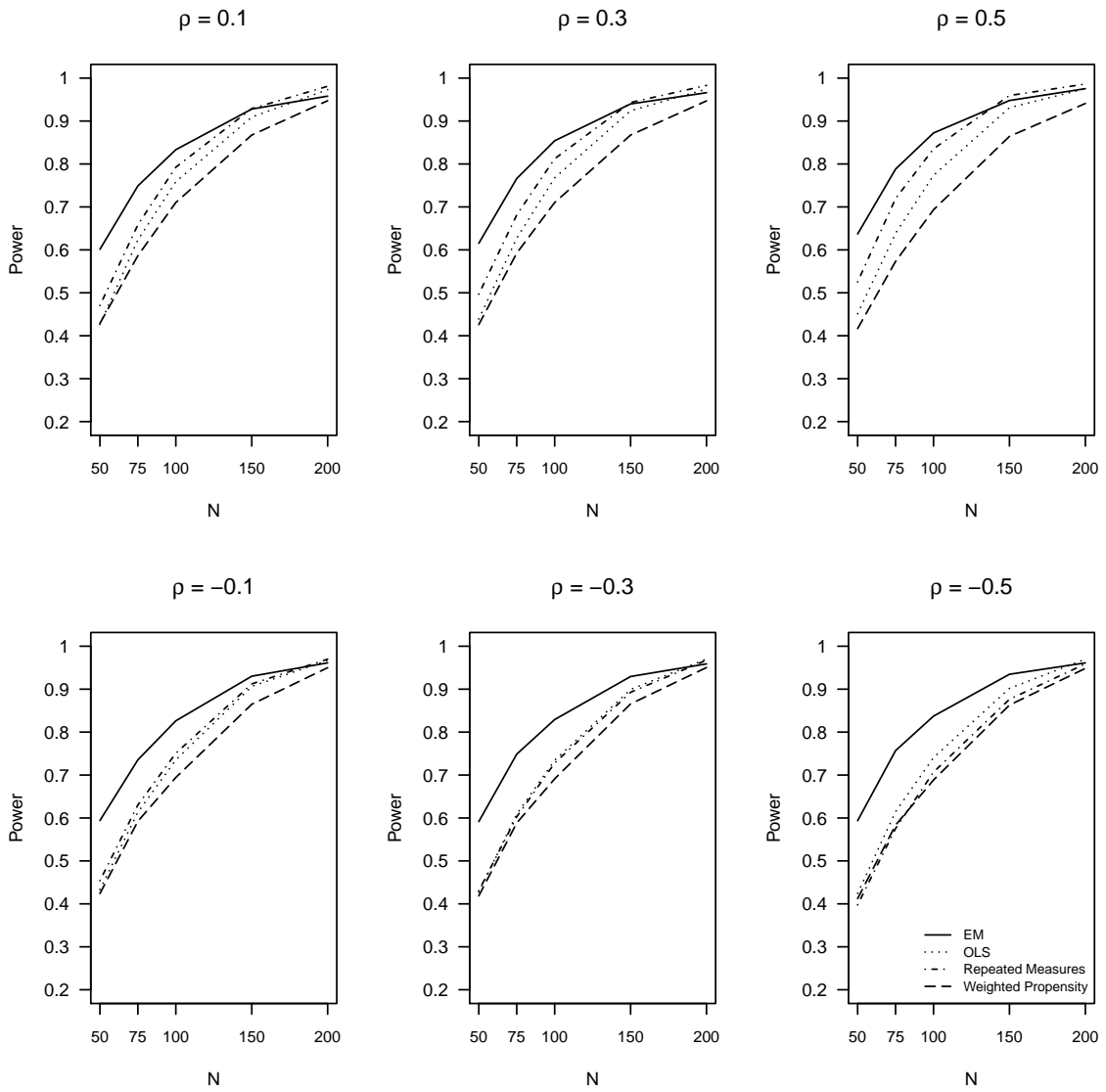


Figure C-3: Power assuming correct response threshold specification - Stage II effect in ‘responders’ set to half of the Stage II effect in ‘non-responders’

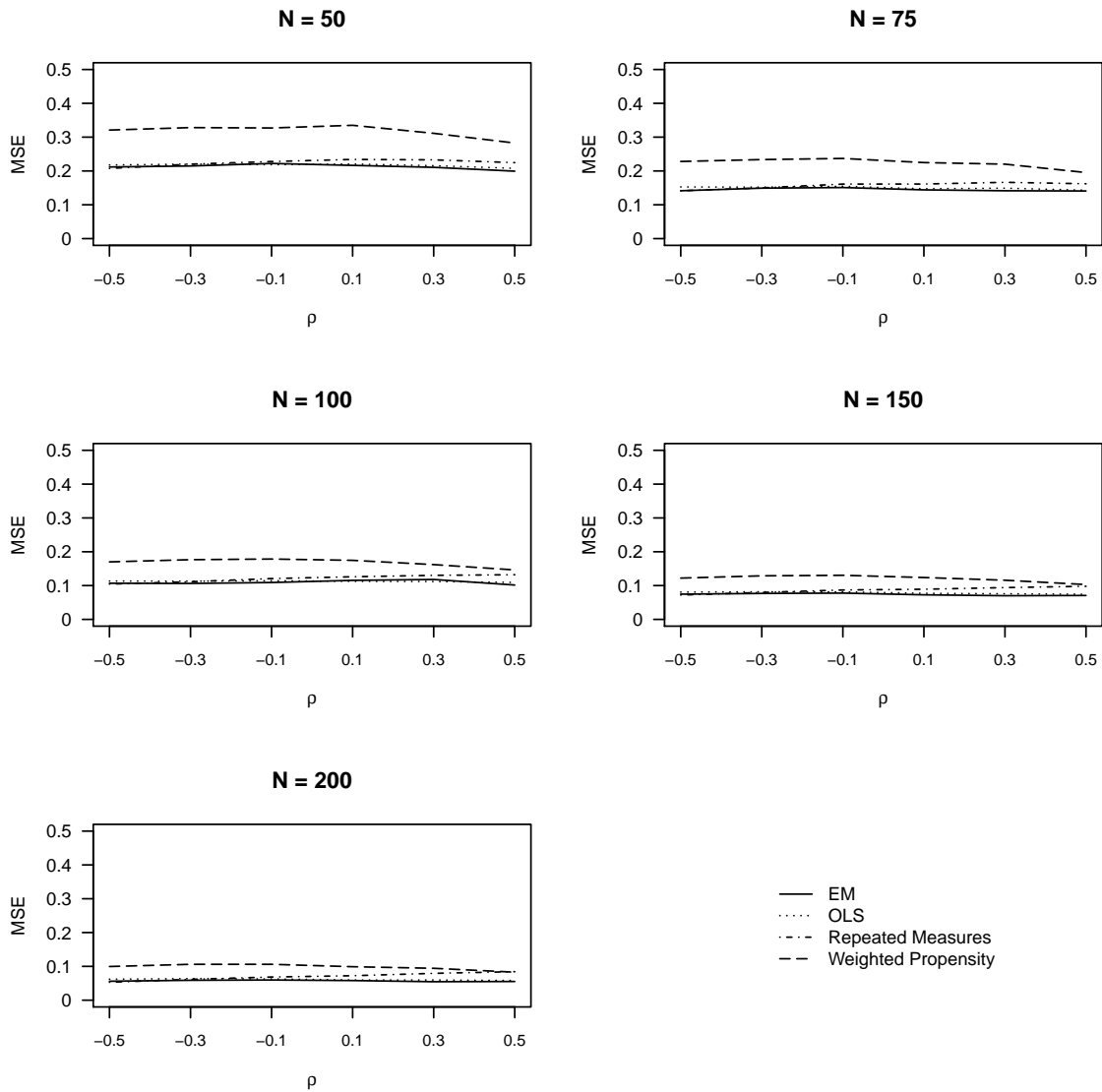


Figure C.4: MSE assuming correct response threshold specification - Stage II effect in ‘responders’ set to two times of the Stage II effect in ‘non-responders’

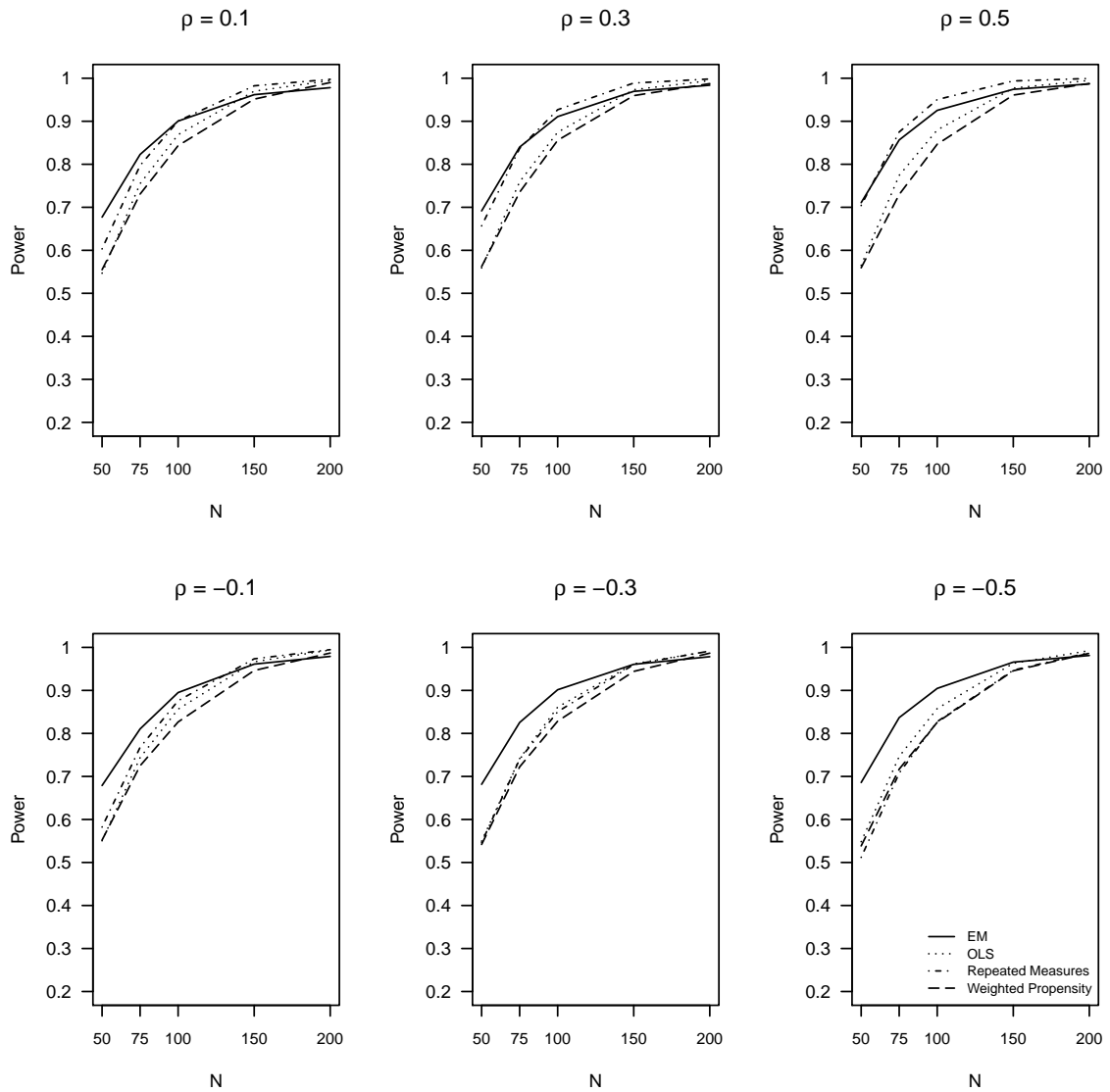


Figure C-5: Power assuming correct response threshold specification - Stage II effect in ‘responders’ set to two times of the Stage II effect in ‘non-responders’

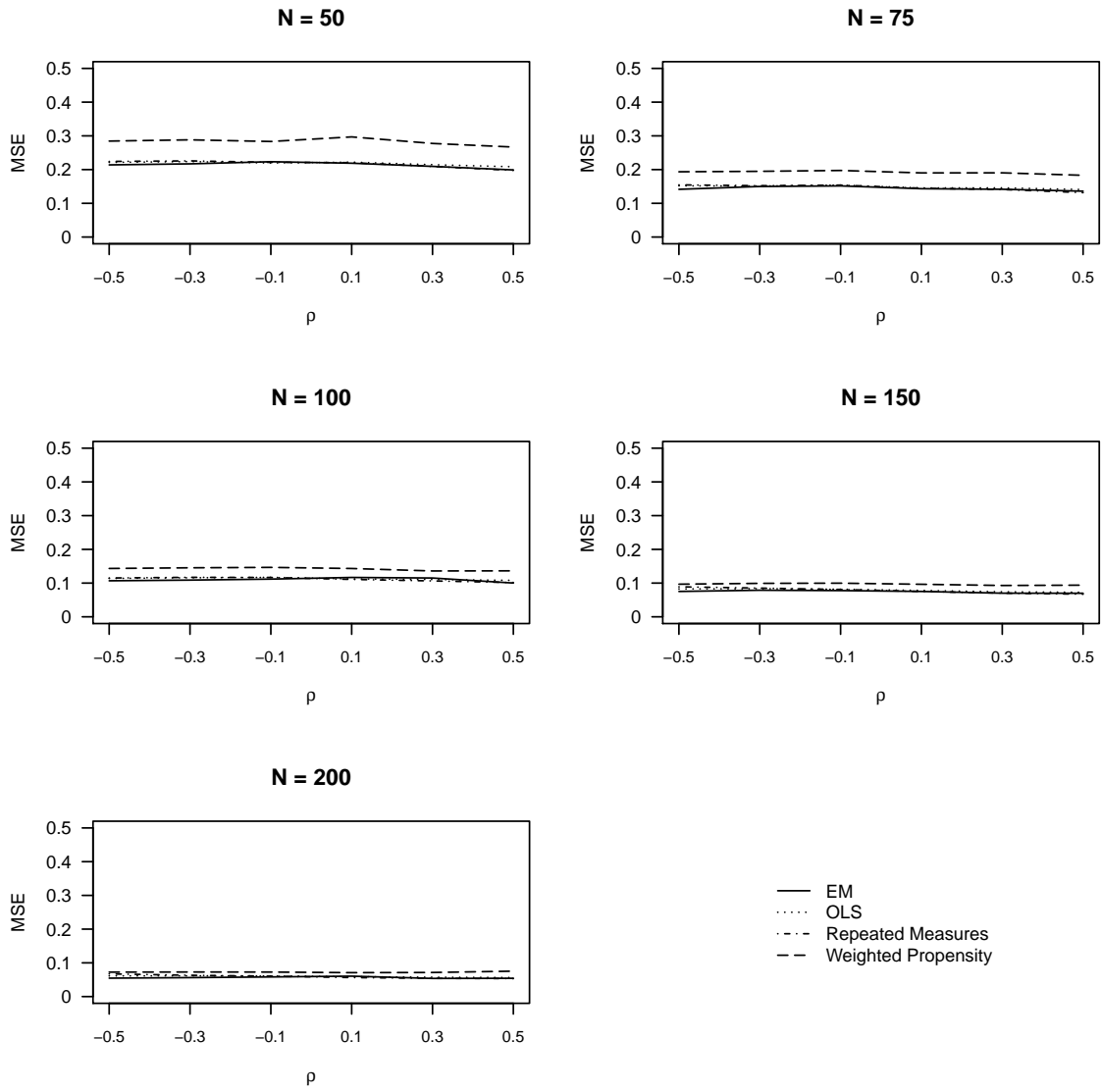


Figure C-6: MSE assuming correct response threshold specification - Stage II effect in 'responders' set to 0

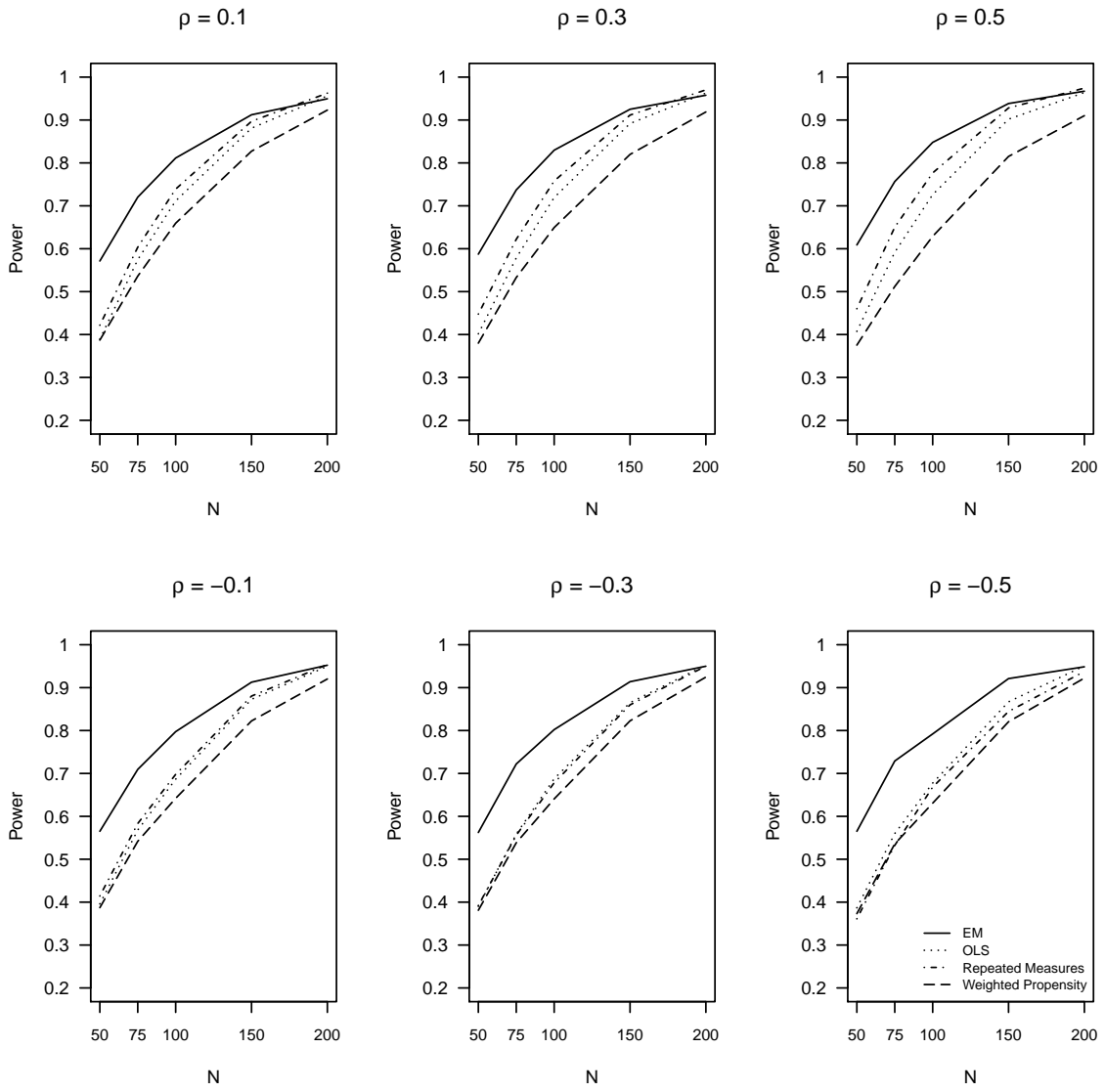


Figure C-7: Power assuming correct response threshold specification - Stage II effect in 'responders' set to 0

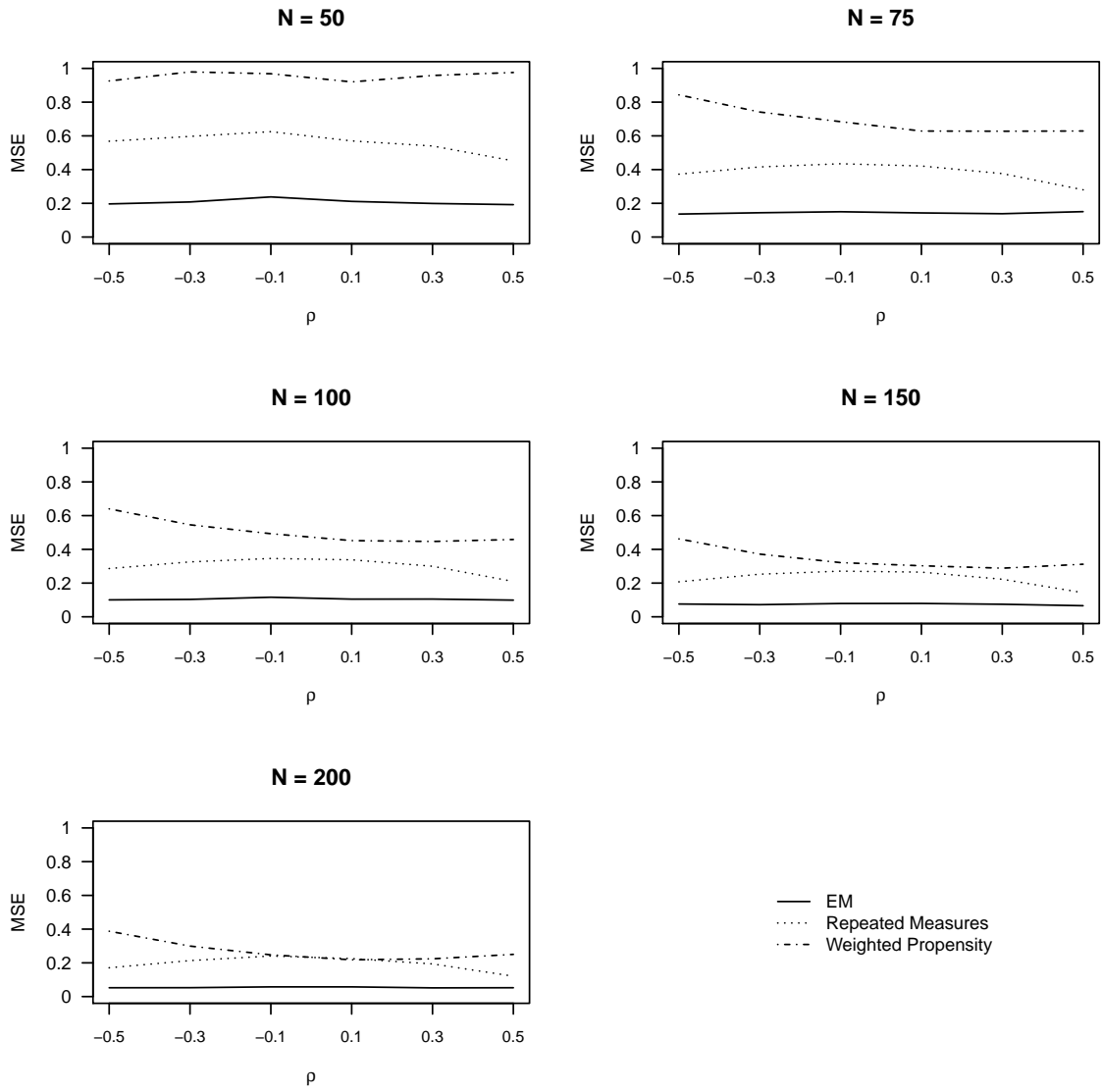


Figure C-8: MSE assuming equal treatment effects but response threshold misspecification

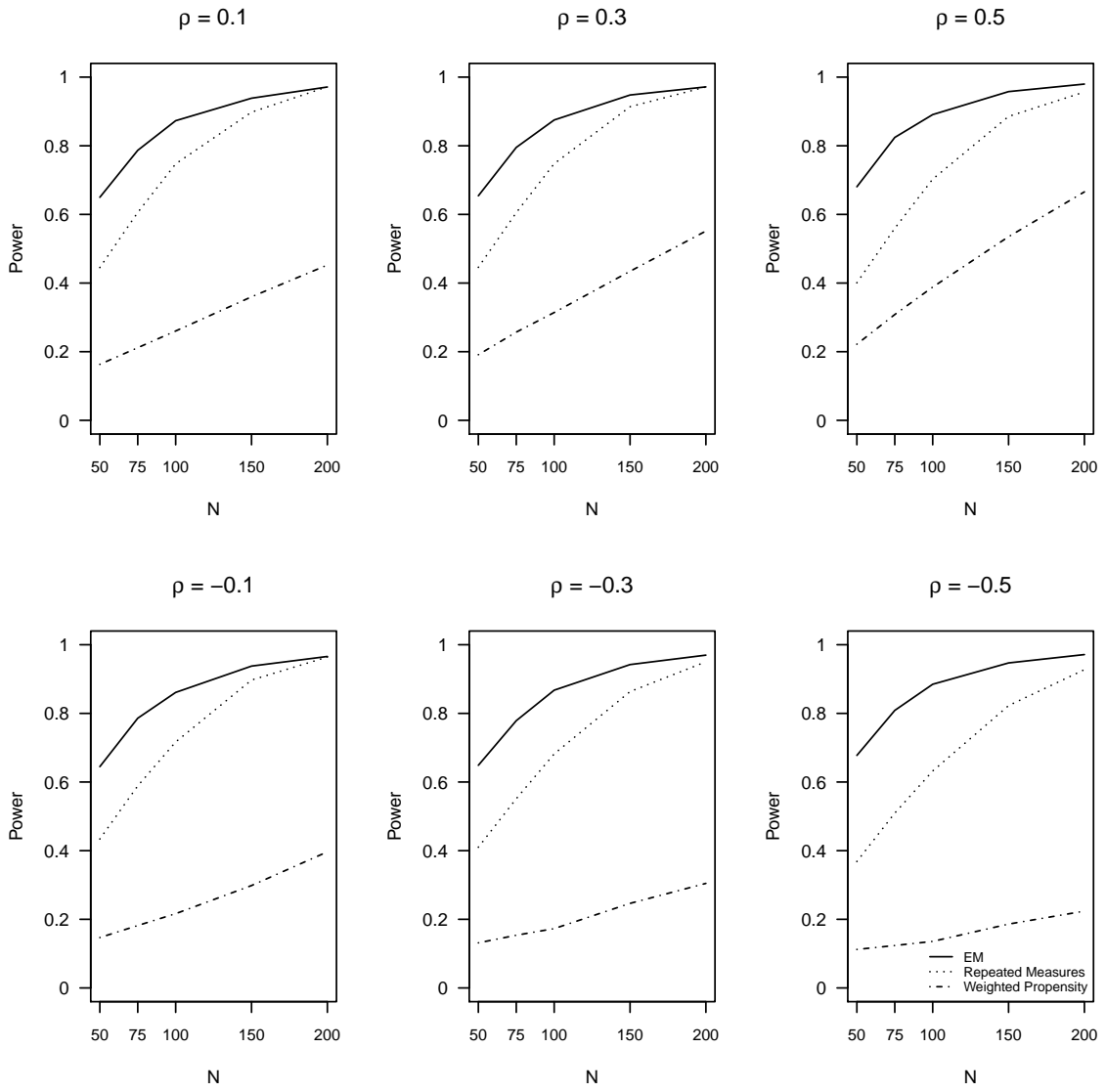


Figure C-9: Power assuming equal treatment effects but response threshold misspecification

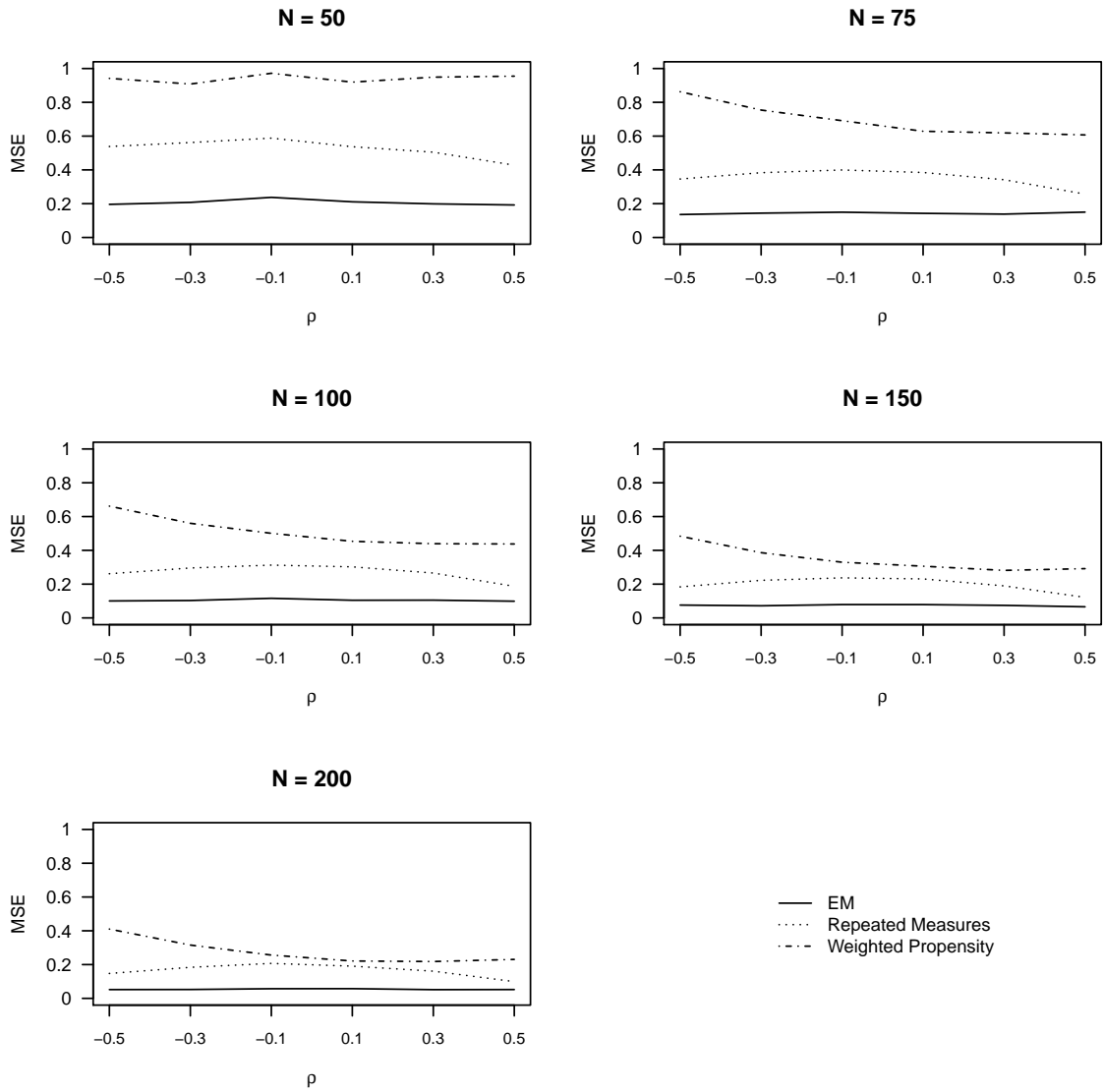


Figure C-10: MSE assuming response threshold misspecification - Stage II effect in 'responders' set to half of the Stage II effect in 'non-responders'

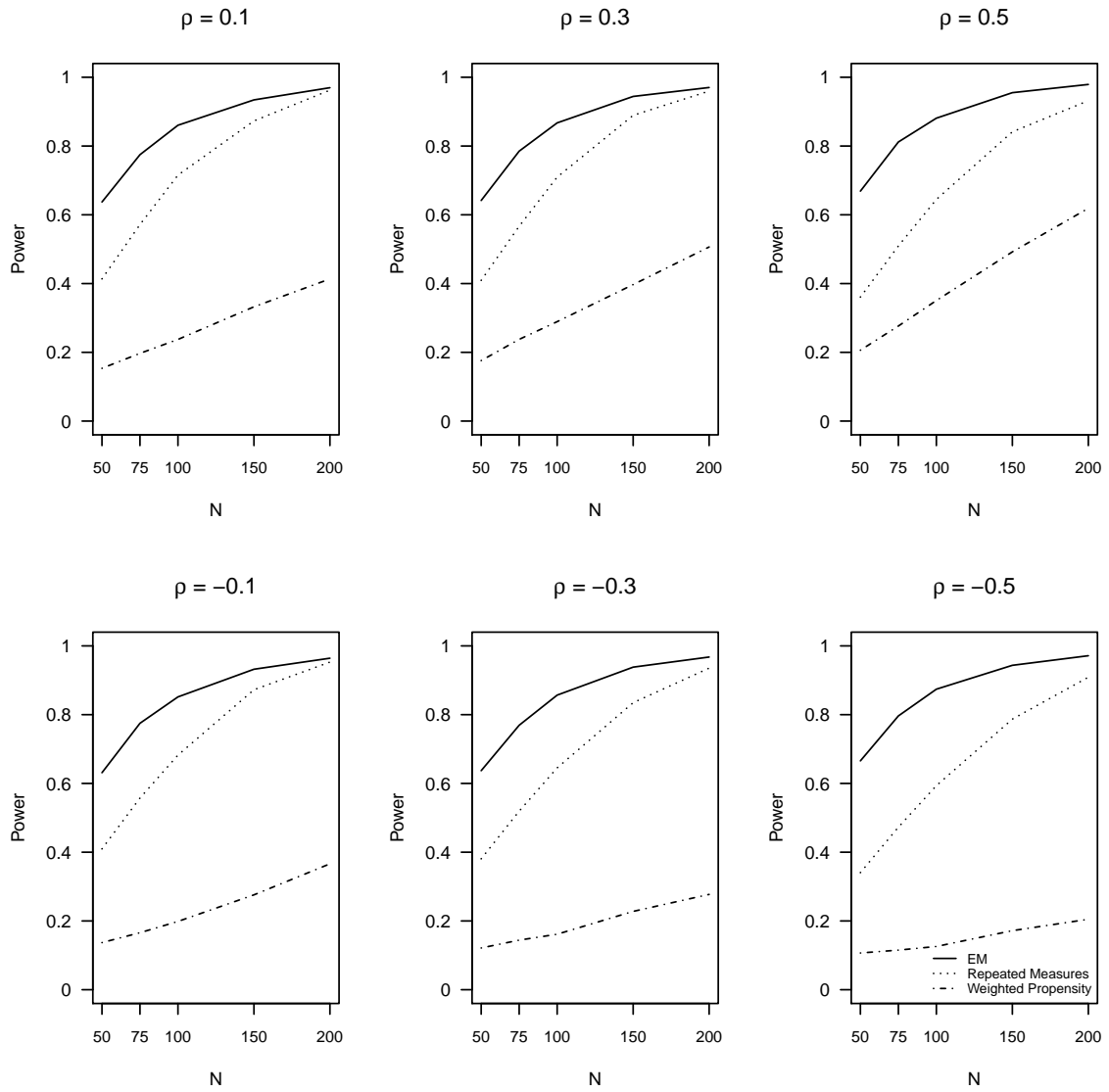


Figure C-11: Power assuming response threshold misspecification - Stage II effect in 'responders' set to half of the Stage II effect in 'non-responders'

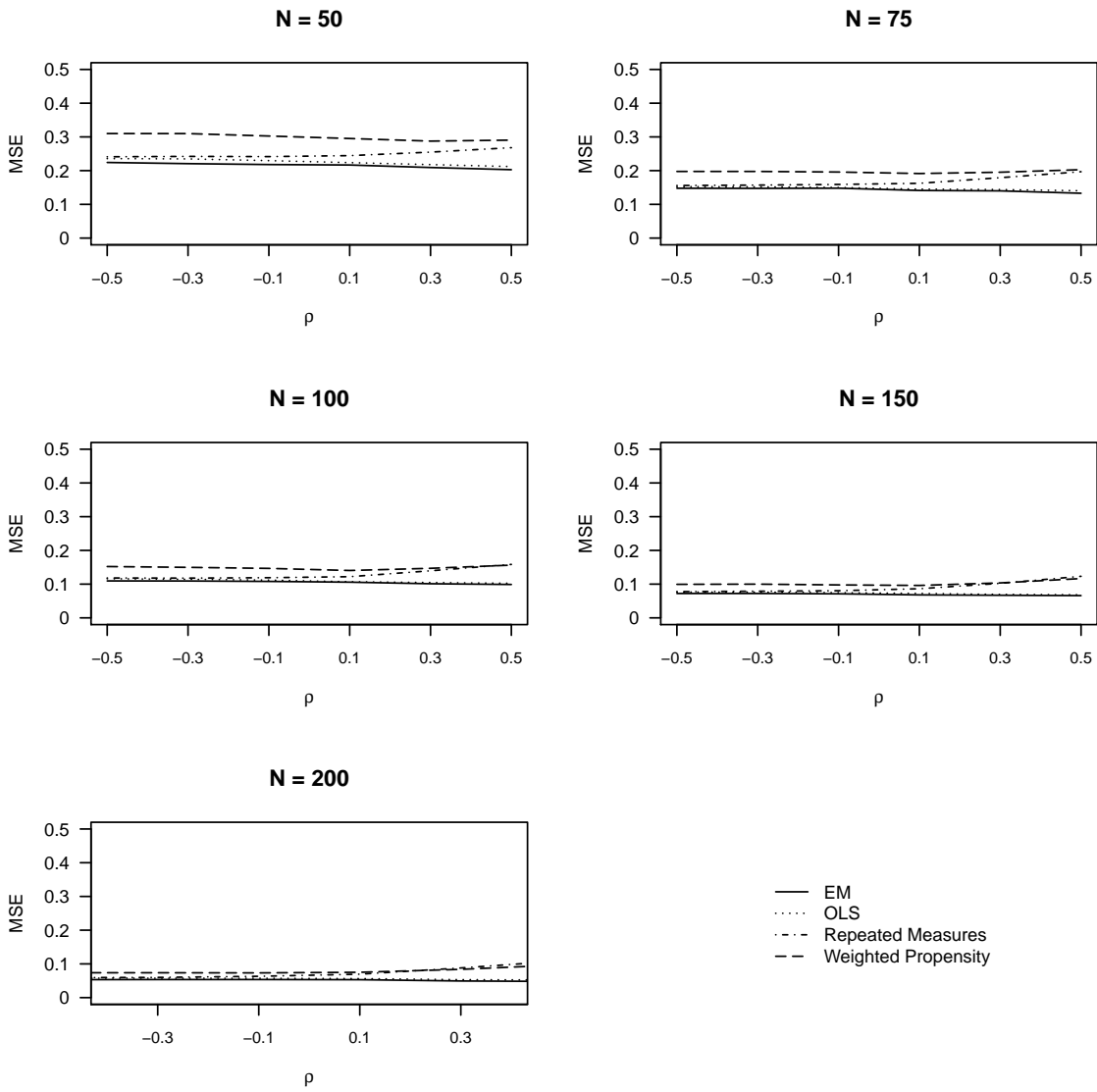


Figure C-12: MSE assuming no ‘placebo-responder’

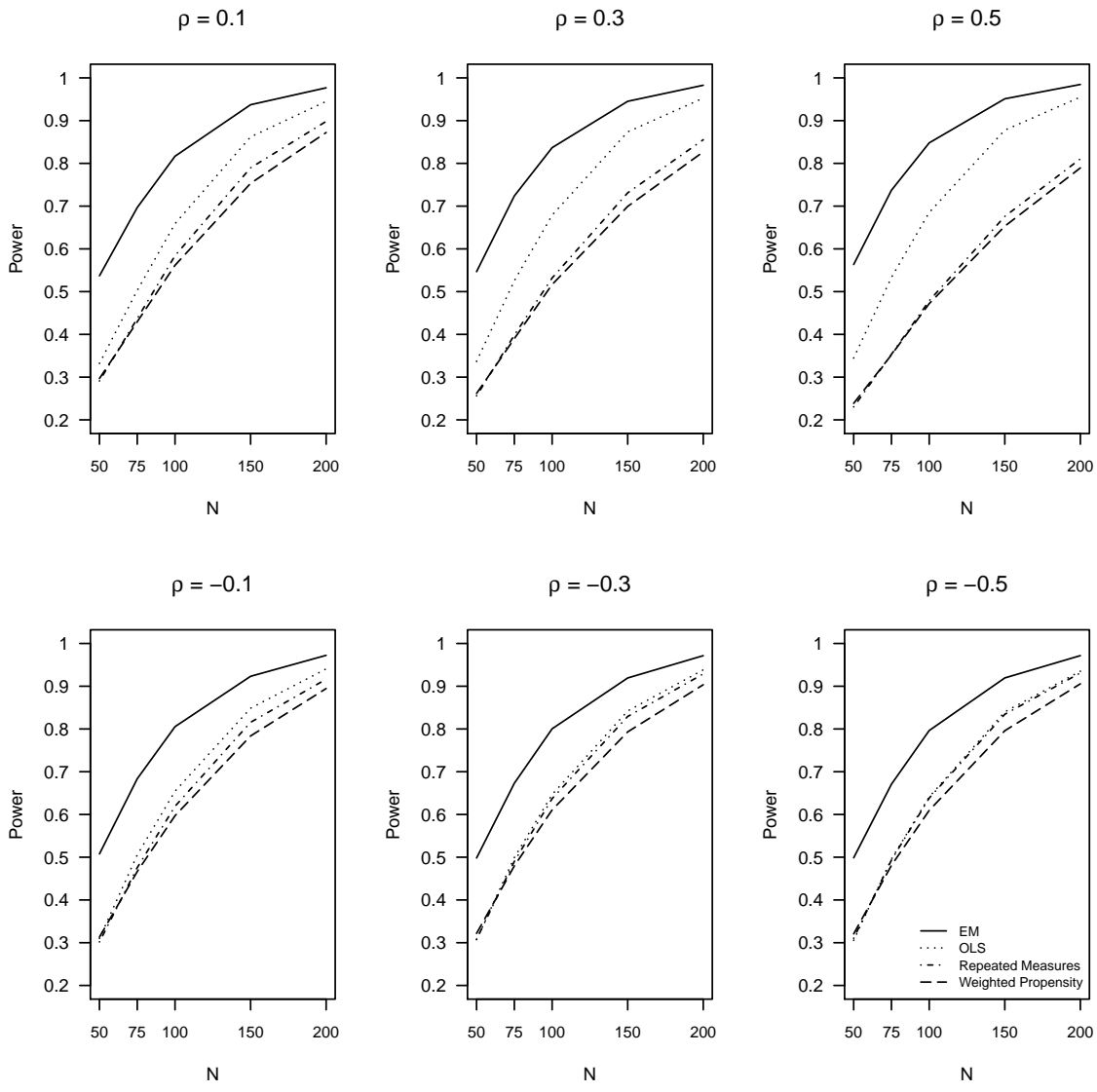


Figure C-13: Power assuming no ‘placebo-responder’

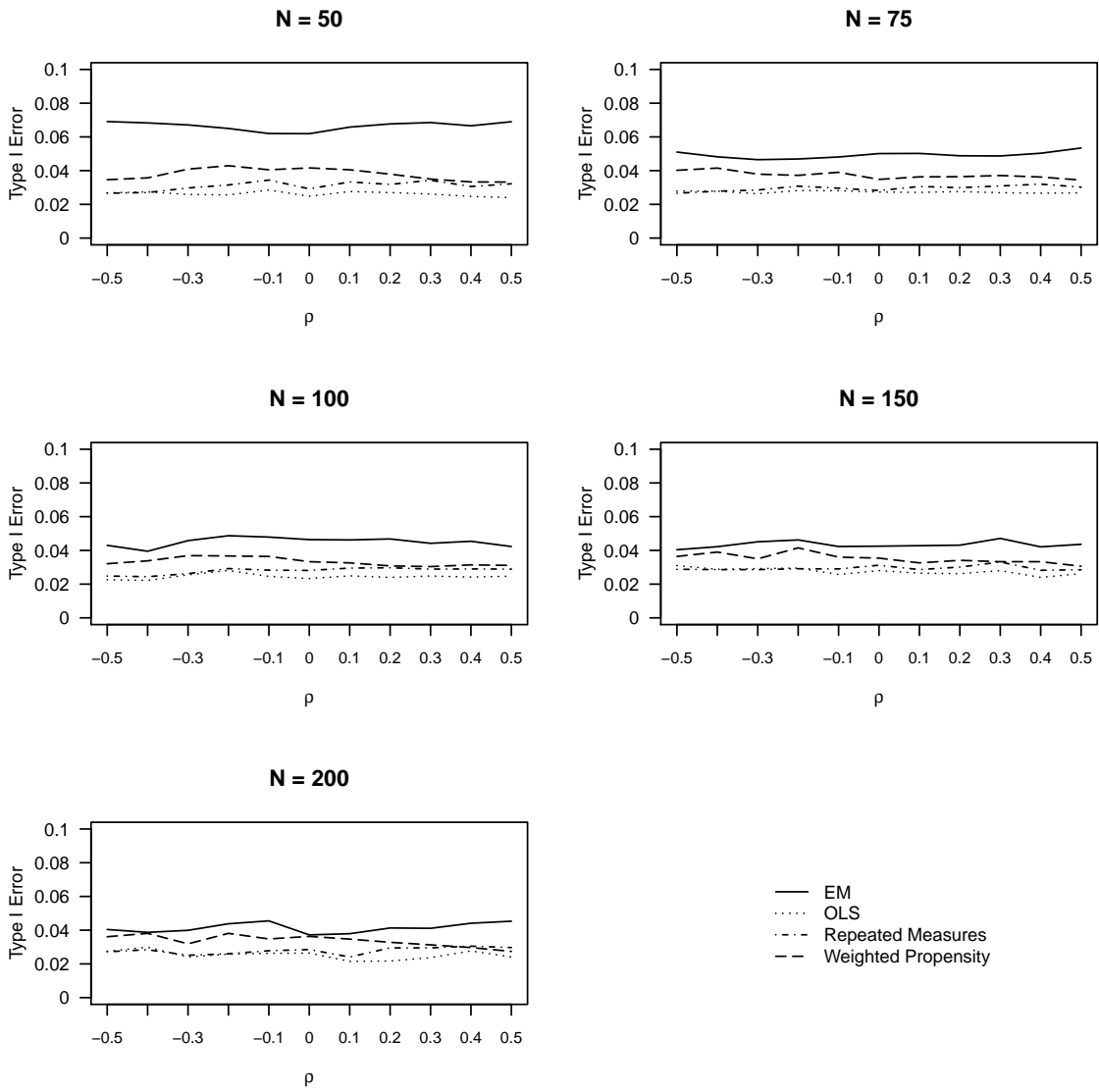


Figure C-14: Type I error with $SD_2=6$

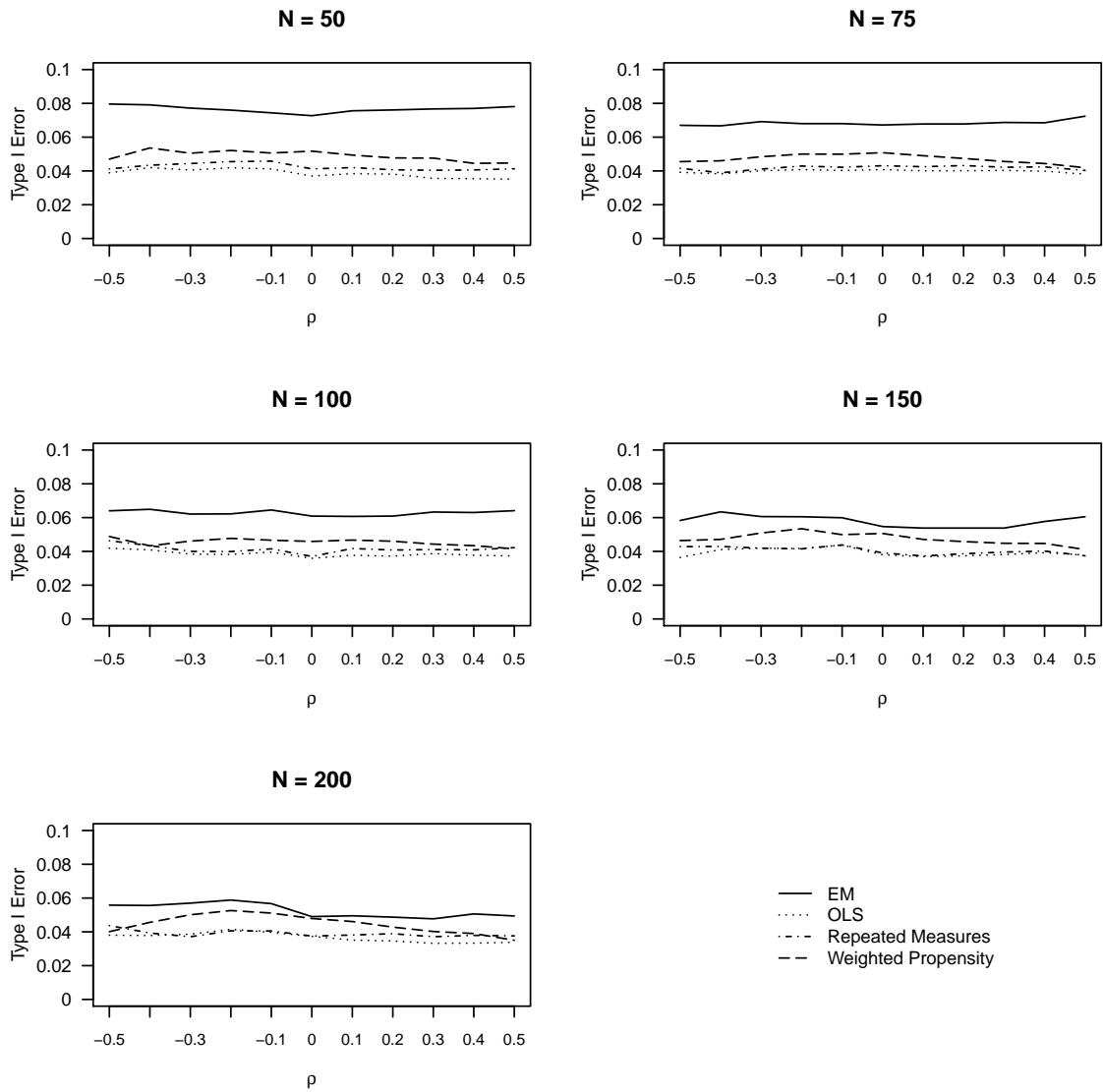


Figure C-15: Type I error with $SD_2=7$

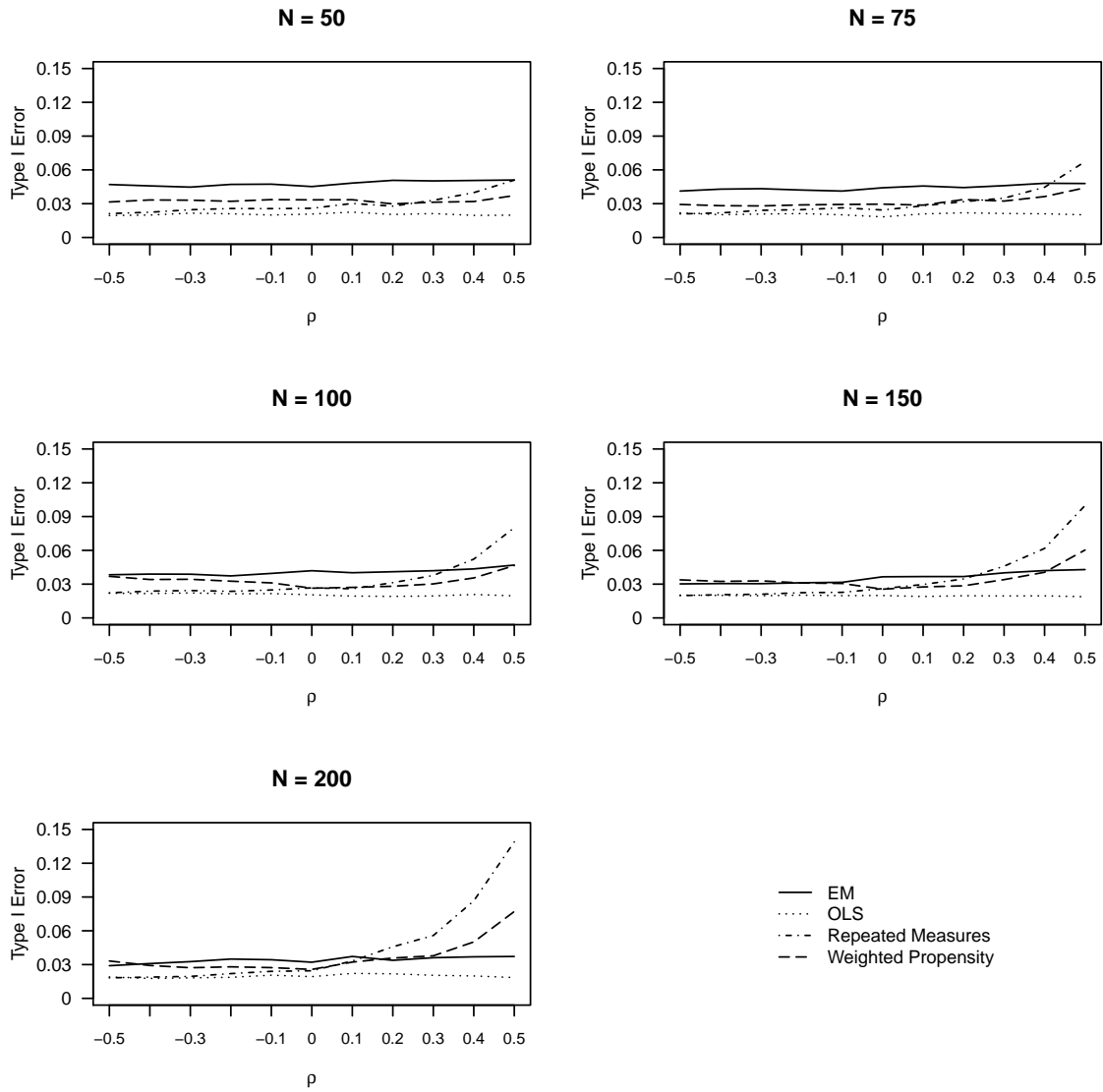


Figure C-16: Type I error with outcome changes in Stage II equal to 80% of that in Stage I

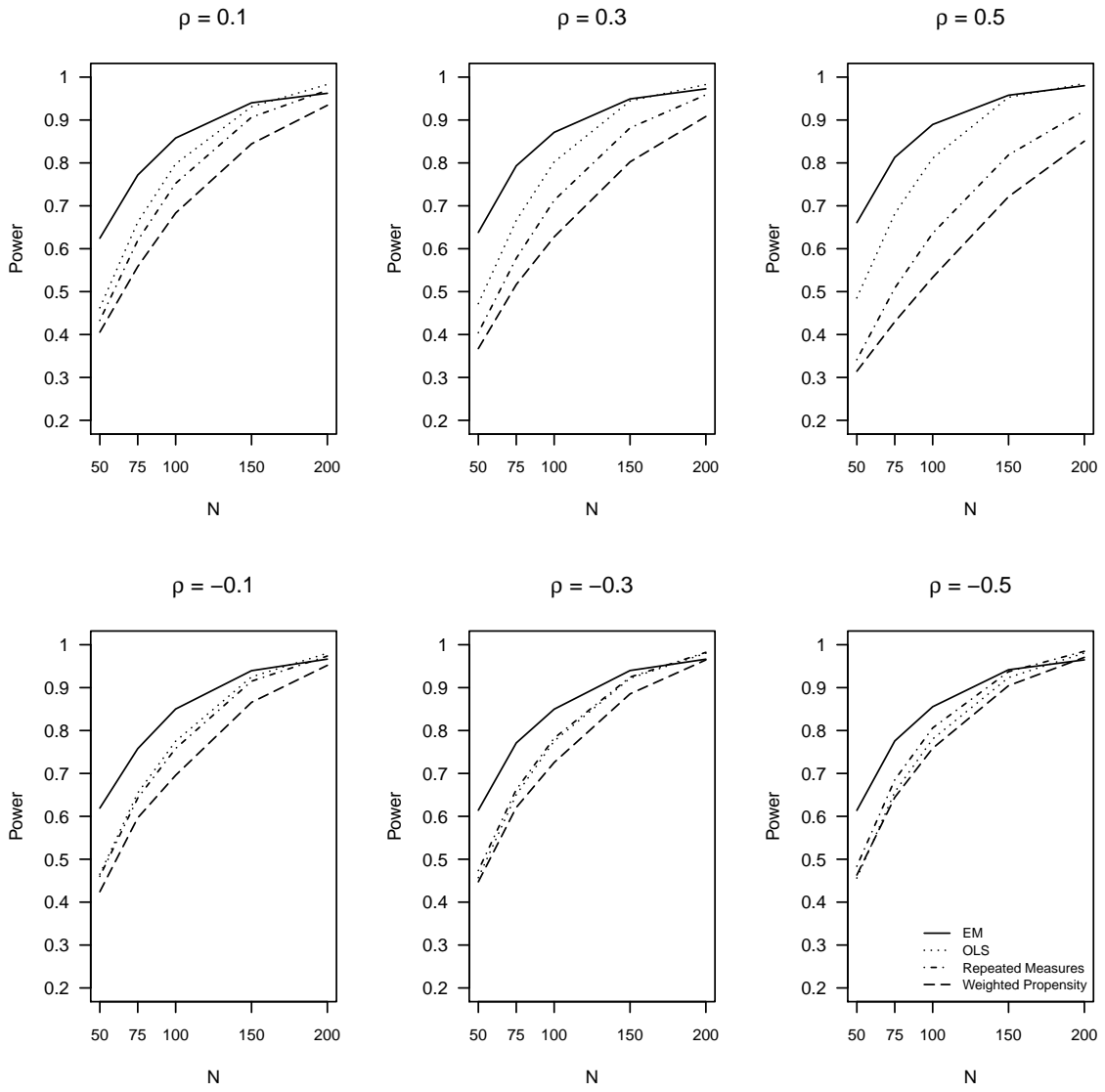


Figure C-17: Power with outcome changes in Stage II equal to 80% of that in Stage I

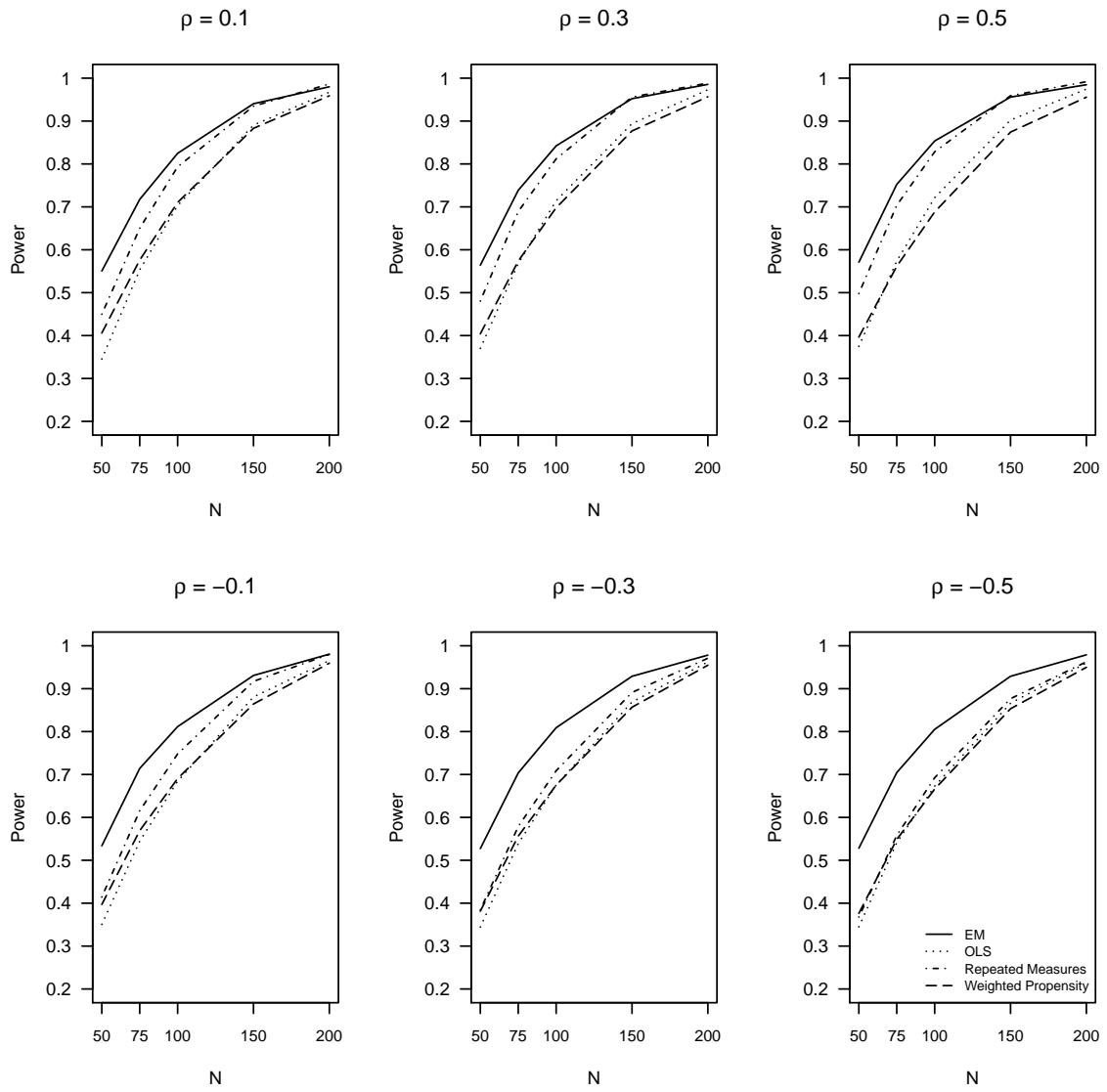


Figure C-18: Power with $\delta_3=9.6$ and $\delta_4=19.3$

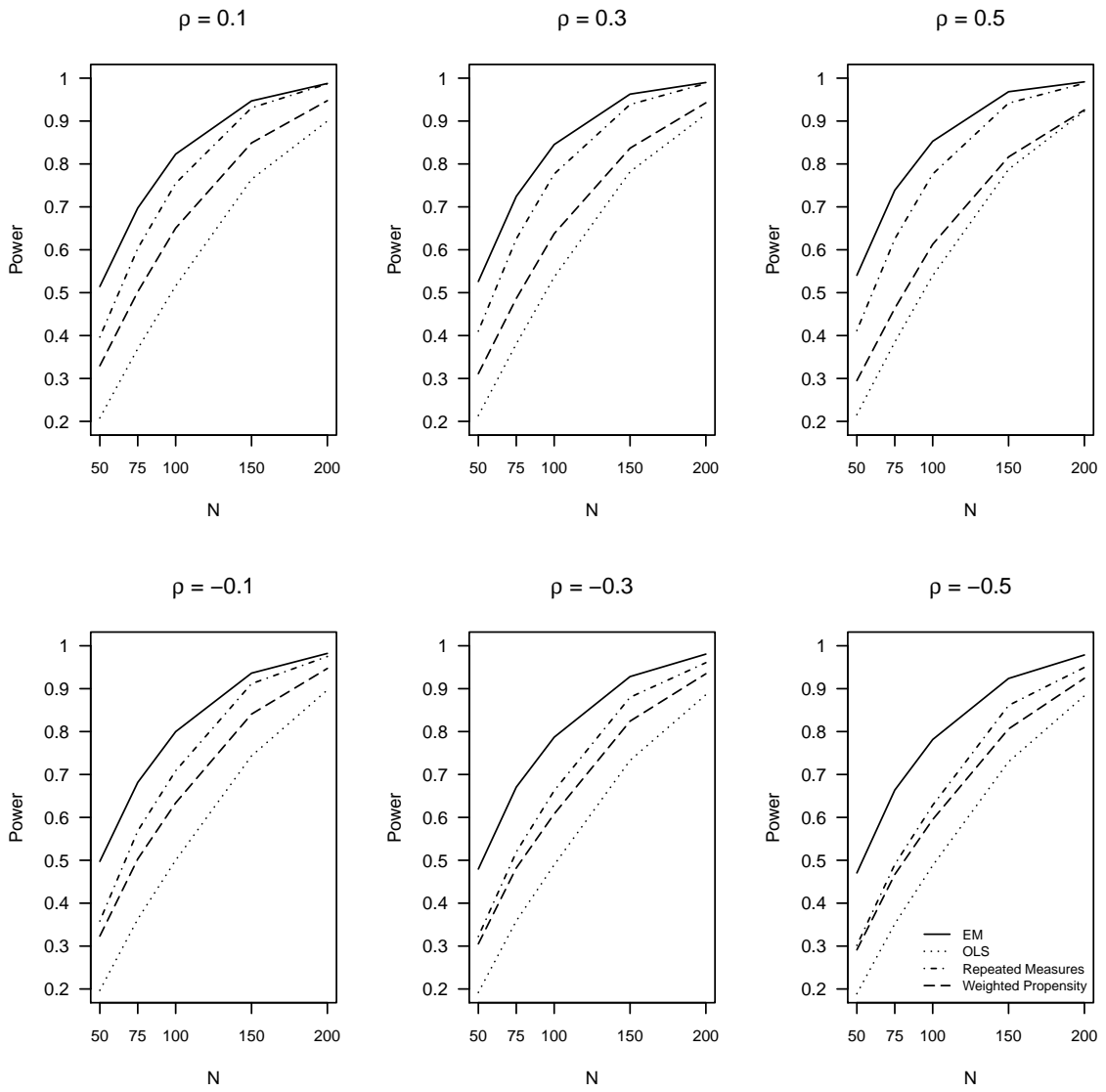


Figure C-19: Power with $\delta_3=8$ and $\delta_4=23.1$

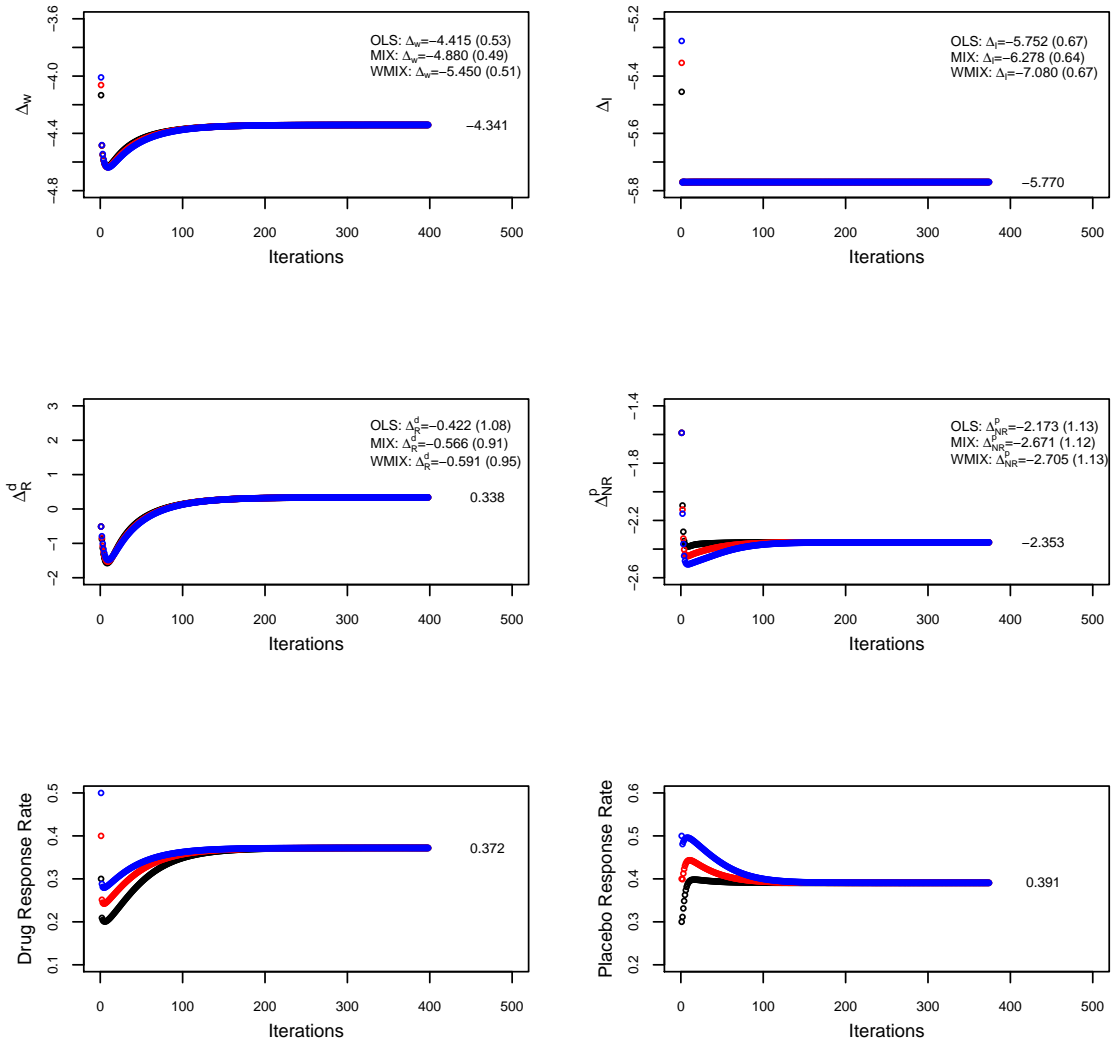


Figure C·20: ADAPT-A trial treatment effect estimates

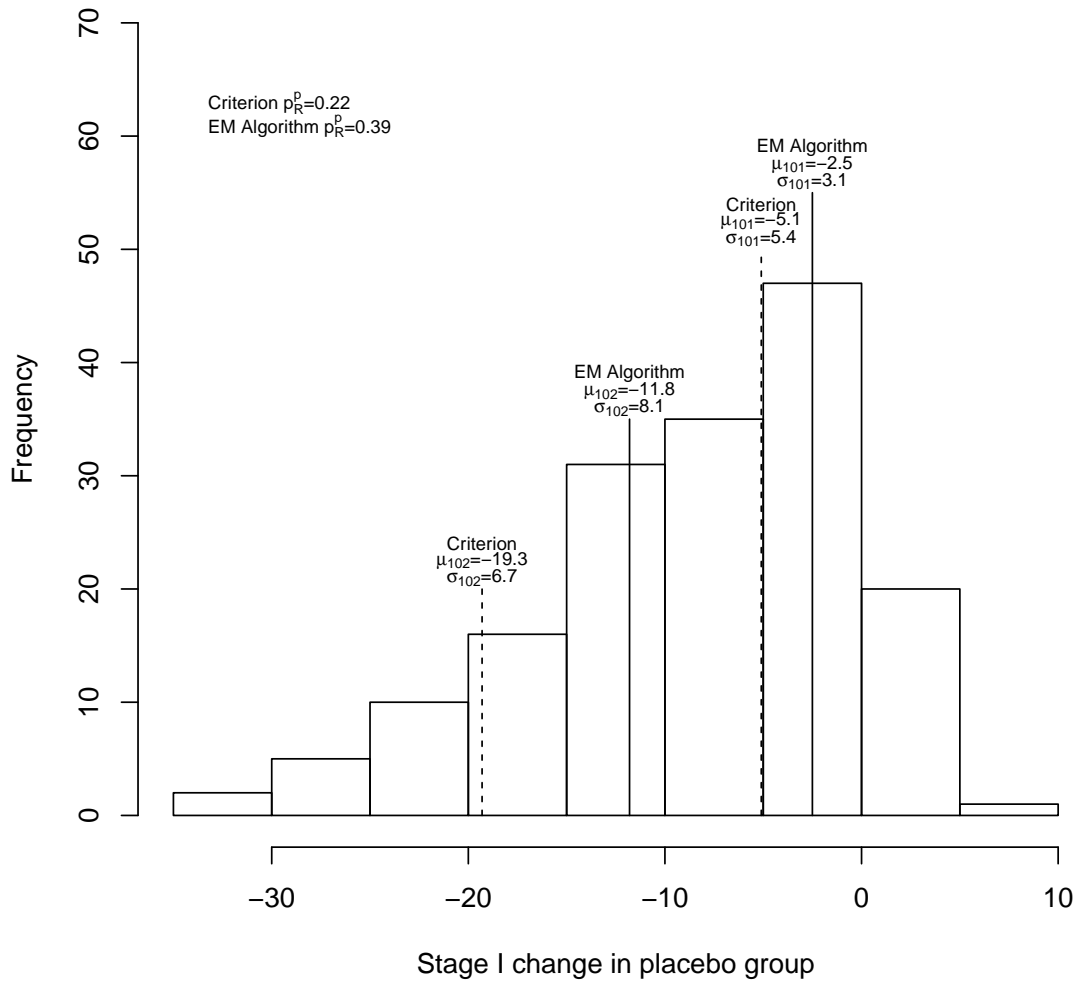


Figure C-21: ADAPT-A trial placebo response classification

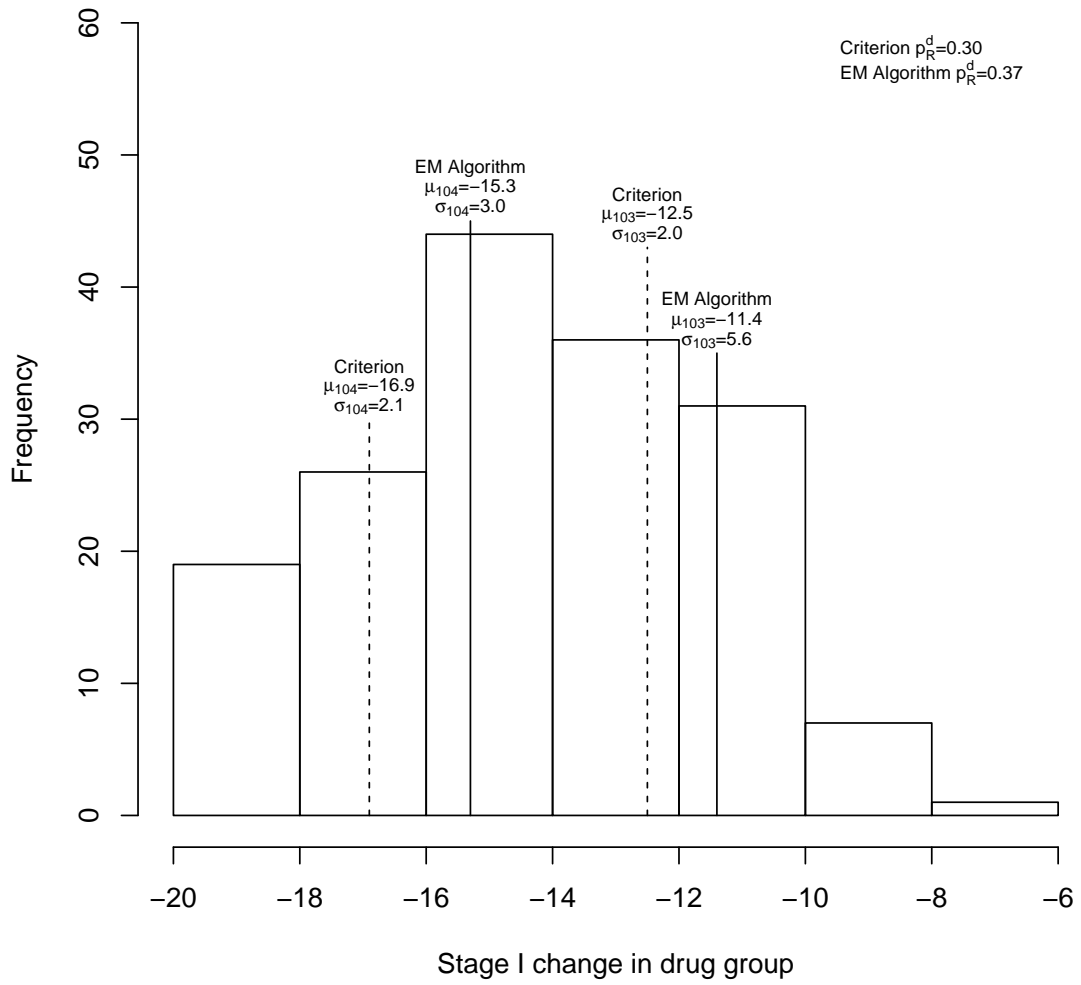


Figure C-22: ADAPT-A trial drug response classification

Appendix D

Parameter Selection in Sequential Enriched Design

Table D.1: Comparison of different designs under Chen's setting

$\mu_R^d = -3.5, \mu_{NR}^d = -3.0, \mu_R^p = -3.0, \mu_{NR}^p = -2.0, \mu_3 = -3.0$																			
Parallel design				Placebo Lead-in				SPCD				TED				SED			
N	p_1	p_2	p_3	p_4	MSE	Power	MSE	Power	MSE	Power	MSE	Power	MSE	Power	MSE	Power	MSE	Power	
50	0	0	0.9	0.1	1.39	0.24	1.37	0.24	1.21	0.30	0.95	0.34	1.16	0.30	1.16	0.30	1.16	0.30	
100	0	0	0.9	0.1	0.71	0.43	0.68	0.42	0.58	0.51	0.45	0.60	0.56	0.51	0.56	0.51	0.56	0.51	
150	0	0	0.9	0.1	0.46	0.59	0.46	0.58	0.37	0.67	0.29	0.77	0.35	0.68	0.35	0.68	0.35	0.68	
200	0	0	0.9	0.1	0.35	0.70	0.35	0.69	0.29	0.79	0.22	0.88	0.27	0.80	0.27	0.80	0.27	0.80	
50	0.1	0	0.8	0.1	1.41	0.22	1.42	0.21	1.22	0.26	0.95	0.31	1.15	0.27	1.15	0.27	1.15	0.27	
100	0.1	0	0.8	0.1	0.73	0.38	0.72	0.37	0.60	0.45	0.47	0.53	0.57	0.47	0.57	0.47	0.57	0.47	
150	0.1	0	0.8	0.1	0.47	0.53	0.49	0.52	0.39	0.60	0.32	0.70	0.38	0.63	0.38	0.63	0.38	0.63	
200	0.1	0	0.8	0.1	0.37	0.64	0.37	0.63	0.30	0.74	0.24	0.82	0.28	0.75	0.28	0.75	0.28	0.75	
50	0.1	0.1	0.7	0.1	1.47	0.19	1.49	0.19	1.29	0.23	1.02	0.25	1.20	0.22	1.20	0.22	1.20	0.22	
100	0.1	0.1	0.7	0.1	0.79	0.32	0.78	0.31	0.64	0.44	0.54	0.44	0.61	0.39	0.61	0.39	0.61	0.39	
150	0.1	0.1	0.7	0.1	0.54	0.43	0.53	0.44	0.46	0.51	0.38	0.59	0.44	0.54	0.44	0.54	0.44	0.54	
200	0.1	0.1	0.7	0.1	0.43	0.54	0.42	0.54	0.36	0.64	0.31	0.72	0.34	0.66	0.34	0.66	0.34	0.66	
50	0.1	0.1	0.6	0.2	1.51	0.18	1.52	0.18	1.32	0.21	1.07	0.24	1.25	0.22	1.25	0.22	1.25	0.22	
100	0.1	0.1	0.6	0.2	0.83	0.30	0.81	0.29	0.68	0.35	0.57	0.42	0.63	0.38	0.63	0.38	0.63	0.38	
150	0.1	0.1	0.6	0.2	0.57	0.40	0.56	0.41	0.49	0.48	0.43	0.56	0.45	0.51	0.45	0.51	0.45	0.51	
200	0.1	0.1	0.6	0.2	0.47	0.50	0.45	0.51	0.39	0.60	0.34	0.69	0.36	0.63	0.36	0.63	0.36	0.63	
50	0.2	0.1	0.5	0.2	1.58	0.16	1.59	0.16	1.42	0.19	1.17	0.22	1.29	0.19	1.29	0.19	1.29	0.19	
100	0.2	0.1	0.5	0.2	0.90	0.26	0.88	0.26	0.78	0.31	0.65	0.36	0.70	0.32	0.70	0.32	0.70	0.32	
150	0.2	0.1	0.5	0.2	0.65	0.35	0.64	0.36	0.58	0.41	0.49	0.50	0.52	0.45	0.52	0.45	0.52	0.45	
200	0.2	0.1	0.5	0.2	0.55	0.44	0.53	0.45	0.48	0.52	0.41	0.62	0.43	0.57	0.43	0.57	0.43	0.57	
50	0.2	0.1	0.4	0.3	1.63	0.15	1.64	0.15	1.47	0.18	1.17	0.20	1.34	0.18	1.34	0.18	1.34	0.18	
100	0.2	0.1	0.4	0.3	0.95	0.24	0.93	0.24	0.83	0.29	0.69	0.33	0.77	0.30	0.77	0.30	0.77	0.30	
150	0.2	0.1	0.4	0.3	0.69	0.32	0.69	0.33	0.63	0.38	0.53	0.46	0.58	0.42	0.58	0.42	0.58	0.42	
200	0.2	0.1	0.4	0.3	0.59	0.41	0.58	0.42	0.53	0.49	0.46	0.58	0.48	0.52	0.48	0.52	0.48	0.52	
50	0.3	0.1	0.4	0.2	1.68	0.14	1.72	0.14	1.51	0.15	1.28	0.18	1.41	0.17	1.41	0.17	1.41	0.17	
100	0.3	0.1	0.4	0.2	1.00	0.22	0.98	0.22	0.88	0.26	0.74	0.30	0.80	0.29	0.80	0.29	0.80	0.29	
150	0.3	0.1	0.4	0.2	0.74	0.30	0.74	0.30	0.68	0.35	0.59	0.42	0.62	0.40	0.62	0.40	0.62	0.40	
200	0.3	0.1	0.4	0.2	0.64	0.37	0.62	0.38	0.57	0.45	0.51	0.53	0.53	0.48	0.53	0.48	0.53	0.48	
50	0.3	0.2	0.4	0.1	1.80	0.12	1.75	0.12	1.67	0.14	1.36	0.16	1.49	0.15	1.49	0.15	1.49	0.15	
100	0.3	0.2	0.4	0.1	1.12	0.19	1.04	0.19	1.00	0.23	0.86	0.25	0.88	0.25	0.88	0.25	0.88	0.25	
150	0.3	0.2	0.4	0.1	0.86	0.25	0.82	0.26	0.80	0.30	0.72	0.35	0.71	0.34	0.71	0.34	0.71	0.34	
200	0.3	0.2	0.4	0.1	0.76	0.31	0.71	0.34	0.69	0.39	0.63	0.43	0.62	0.43	0.62	0.43	0.62	0.43	

Table D.2: Comparison of different designs under the alternative setting 1

$\mu_R^d = -4.0, \mu_{NR}^d = -3.0, \mu_R^p = -3.0, \mu_{NR}^p = -1.0, \mu_3 = -3.0$																	
			Parallel design			Placebo Lead-in			SPCD			TED			SED		
N	p_1	p_2	p_3	p_4	MSE	Power	MSE	Power	MSE	Power	MSE	Power	MSE	Power	MSE	Power	
50	0	0	0.9	0.1	1.40	0.70	1.39	0.69	1.15	0.78	0.92	0.85	1.10	0.79			
100	0	0	0.9	0.1	0.72	0.94	0.69	0.94	0.55	0.98	0.45	0.99	0.54	0.98			
150	0	0	0.9	0.1	0.46	0.99	0.47	0.99	0.37	1.00	0.30	1.00	0.36	1.00			
200	0	0	0.9	0.1	0.36	1.00	0.35	1.00	0.27	1.00	0.22	1.00	0.27	1.00			
50	0.1	0	0.8	0.1	1.48	0.63	1.50	0.63	1.26	0.72	0.99	0.80	1.10	0.76			
100	0.1	0	0.8	0.1	0.80	0.89	0.76	0.90	0.62	0.96	0.53	0.98	0.57	0.97			
150	0.1	0	0.8	0.1	0.54	0.98	0.53	0.98	0.44	0.99	0.38	1.00	0.40	1.00			
200	0.1	0	0.8	0.1	0.44	1.00	0.41	1.00	0.35	1.00	0.30	1.00	0.32	1.00			
50	0.1	0.1	0.7	0.1	1.76	0.53	1.62	0.57	1.50	0.63	1.28	0.71	1.33	0.68			
100	0.1	0.1	0.7	0.1	1.07	0.81	0.94	0.85	0.86	0.91	0.78	0.95	0.76	0.92			
150	0.1	0.1	0.7	0.1	0.81	0.94	0.70	0.96	0.69	0.98	0.64	0.99	0.57	0.99			
200	0.1	0.1	0.7	0.1	0.71	0.98	0.58	0.99	0.57	1.00	0.54	1.00	0.48	1.00			
50	0.1	0.1	0.6	0.2	1.89	0.50	1.74	0.53	1.62	0.59	1.39	0.66	1.47	0.64			
100	0.1	0.1	0.6	0.2	1.20	0.78	1.05	0.82	0.98	0.89	0.90	0.92	0.86	0.91			
150	0.1	0.1	0.6	0.2	0.93	0.92	0.82	0.94	0.81	0.97	0.75	0.99	0.68	0.98			
200	0.1	0.1	0.6	0.2	0.84	0.97	0.70	0.98	0.70	1.00	0.67	1.00	0.59	1.00			
50	0.2	0.1	0.5	0.2	2.21	0.44	2.06	0.48	1.91	0.52	1.69	0.60	1.69	0.58			
100	0.2	0.1	0.5	0.2	1.52	0.71	1.32	0.75	1.29	0.81	1.21	0.88	1.10	0.86			
150	0.2	0.1	0.5	0.2	1.25	0.87	1.07	0.89	1.11	0.94	1.07	0.97	0.89	0.96			
200	0.2	0.1	0.5	0.2	1.15	0.94	0.94	0.96	1.00	0.98	0.97	0.99	0.81	0.99			
50	0.2	0.1	0.4	0.3	2.39	0.41	2.24	0.44	2.10	0.49	1.87	0.56	1.85	0.53			
100	0.2	0.1	0.4	0.3	1.70	0.67	1.50	0.71	1.48	0.77	1.38	0.85	1.27	0.83			
150	0.2	0.1	0.4	0.3	1.44	0.84	1.25	0.87	1.30	0.91	1.25	0.96	1.09	0.95			
200	0.2	0.1	0.4	0.3	1.34	0.92	1.12	0.95	1.19	0.97	1.16	0.99	1.01	0.98			
50	0.3	0.1	0.4	0.2	2.61	0.37	2.36	0.40	2.27	0.45	2.08	0.51	2.02	0.50			
100	0.3	0.1	0.4	0.2	1.92	0.62	1.65	0.67	1.67	0.74	1.59	0.81	1.42	0.80			
150	0.3	0.1	0.4	0.2	1.65	0.80	1.39	0.85	1.48	0.91	1.45	0.94	1.24	0.93			
200	0.3	0.1	0.4	0.2	1.55	0.89	1.27	0.93	1.38	0.95	1.35	0.98	1.14	0.98			
50	0.3	0.2	0.4	0.1	3.09	0.31	2.65	0.36	2.77	0.38	2.50	0.44	2.32	0.45			
100	0.3	0.2	0.4	0.1	2.40	0.53	1.97	0.61	2.10	0.65	2.04	0.72	1.72	0.74			
150	0.3	0.2	0.4	0.1	2.12	0.71	1.73	0.78	1.91	0.82	1.89	0.87	1.53	0.88			
200	0.3	0.2	0.4	0.1	2.03	0.82	1.61	0.88	1.80	0.91	1.79	0.96	1.45	0.96			

Table D.3: Comparison of different designs under the alternative setting 2

$\mu_R^d = -4.0, \mu_{NR}^d = -2.0, \mu_R^p = -2.0, \mu_{NR}^p = -1.0, \mu_3 = -3.0$														
Parallel design			Placebo Lead-in			SPCD			TED			SED		
N	p_1	p_2	p_3	p_4	MSE	Power	MSE	Power	MSE	Power	MSE	Power	MSE	Power
50	0	0	0.9	0.1	1.44	0.66	1.44	0.65	1.19	0.75	0.95	0.82	1.18	0.75
100	0	0	0.9	0.1	0.75	0.92	0.73	0.92	0.58	0.97	0.49	0.98	0.57	0.97
150	0	0	0.9	0.1	0.50	0.99	0.51	0.99	0.40	1.00	0.34	1.00	0.39	1.00
200	0	0	0.9	0.1	0.39	1.00	0.39	1.00	0.30	1.00	0.26	1.00	0.30	1.00
50	0.1	0	0.8	0.1	1.49	0.63	1.48	0.63	1.27	0.72	0.99	0.81	1.16	0.74
100	0.1	0	0.8	0.1	0.80	0.89	0.79	0.90	0.63	0.95	0.51	0.98	0.61	0.96
150	0.1	0	0.8	0.1	0.54	0.98	0.55	0.98	0.44	0.99	0.37	1.00	0.44	0.99
200	0.1	0	0.8	0.1	0.44	1.00	0.43	1.00	0.36	1.00	0.29	1.00	0.34	1.00
50	0.1	0.1	0.7	0.1	1.77	0.53	1.71	0.54	1.54	0.62	1.24	0.71	1.35	0.65
100	0.1	0.1	0.7	0.1	1.08	0.81	1.04	0.82	0.90	0.90	0.78	0.95	0.81	0.92
150	0.1	0.1	0.7	0.1	0.81	0.94	0.79	0.94	0.71	0.98	0.63	0.99	0.63	0.98
200	0.1	0.1	0.7	0.1	0.71	0.98	0.67	0.99	0.62	0.99	0.54	1.00	0.53	1.00
50	0.1	0.1	0.6	0.2	2.06	0.47	1.99	0.47	1.83	0.55	1.52	0.63	1.70	0.58
100	0.1	0.1	0.6	0.2	1.36	0.75	1.32	0.75	1.17	0.84	1.04	0.90	1.09	0.86
150	0.1	0.1	0.6	0.2	1.09	0.90	1.06	0.90	0.98	0.96	0.89	0.98	0.90	0.96
200	0.1	0.1	0.6	0.2	0.99	0.96	0.95	0.96	0.89	0.99	0.80	1.00	0.81	0.99
50	0.2	0.1	0.5	0.2	2.21	0.44	2.16	0.44	1.98	0.52	1.65	0.60	1.80	0.54
100	0.2	0.1	0.5	0.2	1.52	0.71	1.50	0.70	1.36	0.80	1.21	0.88	1.23	0.83
150	0.2	0.1	0.5	0.2	1.25	0.87	1.28	0.86	1.18	0.93	1.06	0.97	1.03	0.95
200	0.2	0.1	0.5	0.2	1.15	0.94	1.15	0.95	1.07	0.98	0.97	0.99	0.93	0.99
50	0.2	0.1	0.4	0.3	2.60	0.37	2.57	0.37	2.39	0.44	2.07	0.52	2.22	0.46
100	0.2	0.1	0.4	0.3	1.92	0.62	1.92	0.61	1.76	0.73	1.59	0.81	1.63	0.75
150	0.2	0.1	0.4	0.3	1.65	0.80	1.69	0.78	1.59	0.88	1.45	0.93	1.45	0.90
200	0.2	0.1	0.4	0.3	1.55	0.89	1.57	0.89	1.48	0.95	1.36	0.98	1.35	0.96
50	0.3	0.1	0.4	0.2	2.40	0.40	2.42	0.39	2.12	0.49	1.85	0.56	2.03	0.50
100	0.3	0.1	0.4	0.2	1.71	0.67	1.74	0.66	1.52	0.77	1.40	0.84	1.44	0.79
150	0.3	0.1	0.4	0.2	1.44	0.84	1.48	0.83	1.34	0.92	1.27	0.95	1.24	0.92
200	0.3	0.1	0.4	0.2	1.34	0.92	1.36	0.92	1.23	0.97	1.16	0.99	1.16	0.98
50	0.3	0.2	0.4	0.1	2.62	0.37	2.51	0.37	2.40	0.45	2.07	0.52	2.22	0.48
100	0.3	0.2	0.4	0.1	1.92	0.63	1.85	0.63	1.74	0.72	1.58	0.82	1.60	0.75
150	0.3	0.2	0.4	0.1	1.65	0.80	1.61	0.80	1.55	0.88	1.42	0.94	1.41	0.90
200	0.3	0.2	0.4	0.1	1.55	0.89	1.49	0.90	1.45	0.95	1.34	0.98	1.30	0.97

Table D.4: Comparison of different designs under the alternative setting 3

		$\mu_R^d = -4.0, \mu_{NR}^d = -3.0, \mu_R^p = -2.0, \mu_{NR}^p = -1.0, \mu_3 = -3.0$														
		Parallel design			Placebo Lead-in			SPCD			TED			SED		
N	p_1	p_2	p_3	p_4	MSE	Power	MSE	Power	MSE	Power	MSE	Power	MSE	Power	MSE	Power
50	0	0	0.9	0.1	1.40	0.70	1.39	0.69	1.15	0.78	0.92	0.85	1.10	0.79	1.10	0.79
100	0	0	0.9	0.1	0.72	0.94	0.69	0.94	0.55	0.98	0.45	0.99	0.54	0.98	0.54	0.98
150	0	0	0.9	0.1	0.46	0.99	0.47	0.99	0.37	1.00	0.30	1.00	0.36	1.00	0.36	1.00
200	0	0	0.9	0.1	0.36	1.00	0.35	1.00	0.27	1.00	0.22	1.00	0.27	1.00	0.27	1.00
50	0.1	0	0.8	0.1	1.43	0.66	1.42	0.66	1.21	0.75	0.93	0.83	1.11	0.77	1.11	0.77
100	0.1	0	0.8	0.1	0.75	0.92	0.73	0.92	0.57	0.97	0.47	0.99	0.55	0.97	0.55	0.97
150	0.1	0	0.8	0.1	0.49	0.99	0.49	0.99	0.39	1.00	0.32	1.00	0.37	1.00	0.37	1.00
200	0.1	0	0.8	0.1	0.39	1.00	0.38	1.00	0.31	1.00	0.25	1.00	0.28	1.00	0.28	1.00
50	0.1	0.1	0.7	0.1	1.55	0.60	1.52	0.61	1.33	0.69	1.07	0.78	1.23	0.72	1.23	0.72
100	0.1	0.1	0.7	0.1	0.87	0.85	0.85	0.88	0.70	0.95	0.60	0.97	0.67	0.95	0.67	0.95
150	0.1	0.1	0.7	0.1	0.61	0.97	0.60	0.97	0.51	0.99	0.45	1.00	0.48	0.99	0.48	0.99
200	0.1	0.1	0.7	0.1	0.51	0.99	0.49	0.99	0.42	1.00	0.36	1.00	0.40	1.00	0.40	1.00
50	0.1	0.1	0.6	0.2	1.64	0.57	1.61	0.58	1.42	0.65	1.13	0.75	1.35	0.68	1.35	0.68
100	0.1	0.1	0.6	0.2	0.96	0.85	0.93	0.85	0.78	0.93	0.67	0.96	0.74	0.93	0.74	0.93
150	0.1	0.1	0.6	0.2	0.70	0.96	0.69	0.96	0.60	0.99	0.53	1.00	0.57	0.99	0.57	0.99
200	0.1	0.1	0.6	0.2	0.59	0.99	0.57	0.99	0.51	1.00	0.45	1.00	0.48	1.00	0.48	1.00
50	0.2	0.1	0.5	0.2	1.74	0.54	1.71	0.54	1.51	0.63	1.24	0.71	1.43	0.64	1.43	0.64
100	0.2	0.1	0.5	0.2	1.06	0.82	1.05	0.82	0.90	0.90	0.78	0.95	0.84	0.91	0.84	0.91
150	0.2	0.1	0.5	0.2	0.80	0.94	0.82	0.95	0.73	0.98	0.63	0.99	0.68	0.98	0.68	0.98
200	0.2	0.1	0.5	0.2	0.70	0.98	0.70	0.98	0.62	0.99	0.54	1.00	0.58	1.00	0.58	1.00
50	0.2	0.1	0.4	0.3	1.87	0.51	1.84	0.50	1.64	0.59	1.37	0.67	1.58	0.60	1.58	0.60
100	0.2	0.1	0.4	0.3	1.19	0.79	1.18	0.79	1.03	0.87	0.91	0.93	0.98	0.89	0.98	0.89
150	0.2	0.1	0.4	0.3	0.93	0.92	0.96	0.93	0.86	0.97	0.77	0.99	0.81	0.97	0.81	0.97
200	0.2	0.1	0.4	0.3	0.83	0.97	0.84	0.97	0.75	0.99	0.69	1.00	0.71	0.99	0.71	0.99
50	0.3	0.1	0.4	0.2	1.87	0.51	1.87	0.49	1.60	0.60	1.42	0.67	1.49	0.61	1.49	0.61
100	0.3	0.1	0.4	0.2	1.19	0.78	1.21	0.79	1.01	0.88	0.94	0.92	0.94	0.90	0.94	0.90
150	0.3	0.1	0.4	0.2	0.93	0.92	0.96	0.93	0.83	0.97	0.78	0.99	0.78	0.97	0.78	0.97
200	0.3	0.1	0.4	0.2	0.83	0.97	0.84	0.97	0.73	0.99	0.69	1.00	0.70	0.99	0.70	0.99
50	0.3	0.2	0.4	0.1	2.02	0.48	1.95	0.47	1.82	0.56	1.54	0.63	1.67	0.59	1.67	0.59
100	0.3	0.2	0.4	0.1	1.34	0.75	1.30	0.76	1.18	0.84	1.06	0.90	1.11	0.86	1.11	0.86
150	0.3	0.2	0.4	0.1	1.08	0.90	1.06	0.90	0.99	0.96	0.91	0.98	0.91	0.96	0.91	0.96
200	0.3	0.2	0.4	0.1	0.98	0.96	0.94	0.97	0.89	0.99	0.82	1.00	0.82	0.99	0.82	0.99

Appendix E

Sample Size Calculation in Sequential Enriched Design

Table E.1: Sample size at enrollment under Chen's setting
$$\mu_R^d = -3.5, \mu_{NR}^d = -3.0, \mu_R^p = -3.0, \mu_{NR}^p = -2.0$$

p1	p2	p3	p4	Analysis sample size	Enrolled sample size	Percentage used in analysis
0	0	0.9	0.1	200	365	54.72
0.1	0	0.8	0.1	200	372	53.76
0.1	0.1	0.7	0.1	200	378	52.84
0.1	0.1	0.6	0.2	200	378	52.84
0.2	0.1	0.5	0.2	200	385	51.88
0.2	0.1	0.4	0.3	200	385	51.88
0.3	0.1	0.4	0.2	200	392	51.02
0.3	0.2	0.4	0.1	200	400	50.00

Table E.2: Sample size at enrollment under the alternative setting 1
$$\mu_R^d = -4.0, \mu_{NR}^d = -3.0, \mu_R^p = -3.0, \mu_{NR}^p = -1.0$$

p1	p2	p3	p4	Analysis sample size	Enrolled sample size	Percentage used in analysis
0	0	0.9	0.1	200	313	63.90
0.1	0	0.8	0.1	200	322	62.02
0.1	0.1	0.7	0.1	200	332	60.15
0.1	0.1	0.6	0.2	200	332	60.15
0.2	0.1	0.5	0.2	200	343	58.31
0.2	0.1	0.4	0.3	200	343	58.31
0.3	0.1	0.4	0.2	200	354	56.50
0.3	0.2	0.4	0.1	200	366	54.64

Analysis sample is the sample that will be used for estimation of drug efficacy. It only includes 'placebo non-responders' selected at the end of Stage 0.

Enrolled sample is the sample that will be screened and enrolled into the study. It includes both 'placebo responders' and 'placebo non-responders' in Stage 0.

Percentage used in analysis is the percentage of the analysis sample among the enrolled sample.

Table E.3: Sample size at enrollment under the alternative setting 2
$$\mu_R^d = -4.0, \mu_{NR}^d = -2.0, \mu_R^p = -2.0, \mu_{NR}^p = -1.0$$

p1	p2	p3	p4	Analysis sample size	Enrolled sample size	Percentage used in analysis
0	0	0.9	0.1	200	313	63.90
0.1	0	0.8	0.1	200	318	62.99
0.1	0.1	0.7	0.1	200	322	62.02
0.1	0.1	0.6	0.2	200	322	62.02
0.2	0.1	0.5	0.2	200	327	61.16
0.2	0.1	0.4	0.3	200	327	61.16
0.3	0.1	0.4	0.2	200	332	60.24
0.3	0.2	0.4	0.1	200	337	59.35

Analysis sample is the sample that will be used for estimation of drug efficacy. It only includes ‘placebo non-responders’ selected at the end of Stage 0.

Enrolled sample is the sample that will be screened and enrolled into the study. It includes both ‘placebo responders’ and ‘placebo non-responders’ in Stage 0.

Percentage used in analysis is the percentage of the analysis sample among the enrolled sample.

Bibliography

- [1] Larry Alphas, Fabrizio Benedetti, W Wolfgang Fleischhacker, and John M Kane. Placebo-related effects in clinical trials in schizophrenia: what is driving this phenomenon and what can be done to minimize it? *International Journal of Neuropsychopharmacology*, 15(7):1003–1014, 2012.
- [2] Lee Baer and Anastasia Ivanova. When should the sequential parallel comparison design be used in clinical trials? *Clinical Investigation*, 3(9):823–833, 2013.
- [3] Brandeis University. Despite great need, pool of new, innovative psychotropic drugs is running dry. <https://heller.brandeis.edu/news/items/releases/2015/drug-pipeline.html>, 2015. Accessed: 2019-08-07.
- [4] Yeh-Fong Chen, Yang Yang, HM James Hung, and Sue-Jane Wang. Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. *Contemporary clinical trials*, 32(4):592–604, 2011.
- [5] Yeh-Fong Chen, Xiangmin Zhang, Roy N Tamura, and Chiung M Chen. A sequential enriched design for target patient population in psychiatric clinical trials. *Statistics in Medicine*, 33(17):2953–2967, 2014.
- [6] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.
- [7] Gheorghe Doros, Pilar Lim, and Yuyin Liu. Addressing high placebo response in neuroscience clinical trials. In *Biopharmaceutical Applied Statistics Symposium*, pages 171–203. Springer, 2018.
- [8] Gheorghe Doros, Michael J Pencina, Denis Rybin, Allison Meisner, and Maurizio Fava. A repeated measures model for analysis of continuous outcomes in sequential parallel comparison design studies. *Statistics in Medicine*, 32(16):2767–2789, 2013.
- [9] Maurizio Fava, A Eden Evins, David J Dorer, and David A Schoenfeld. The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics*, 72:115–127, 2003.
- [10] Maurizio Fava, David Mischoulon, Dan Iosifescu, Janet Witte, Michael Pencina, Martina Flynn, Linda Harper, Michael Levy, Karl Rickels, and Mark Pollack. A double-blind, placebo-controlled study of aripiprazole adjunctive to antidepressant therapy among depressed outpatients with inadequate response to prior antidepressant therapy (adapt-a study). *Psychotherapy and psychosomatics*, 81(2):87–97, 2012.

- [11] Xiaohong Huang and Roy N Tamura. Comparison of test statistics for the sequential parallel design. *Statistics in Biopharmaceutical Research*, 2(1):42–50, 2010.
- [12] Orest Hurko and John L Ryan. Translational research in central nervous system drug discovery. *NeuroRx*, 2(4):671–682, 2005.
- [13] Anastasia Ivanova, Bahjat Qaqish, and David A Schoenfeld. Optimality, sample size, and power calculations for the sequential parallel comparison design. *Statistics in medicine*, 30(23):2793–2803, 2011.
- [14] Anastasia Ivanova and Roy N Tamura. A two-way enriched clinical trial design: combining advantages of placebo lead-in and randomized withdrawal. *Statistical Methods in Medical Research*, 24(6):871–890, 2015.
- [15] Anastasia Ivanova, Zhiwei Zhang, Laura Thompson, Ying Yang, Richard M Kotz, and Xin Fang. Can sequential parallel comparison design and two-way enriched design be useful in medical device clinical trials? *Journal of biopharmaceutical statistics*, 26(1):167–177, 2016.
- [16] Aaron S Kemp, Nina R Schooler, Amir H Kalali, Larry Alphs, Ravi Anand, George Awad, Michael Davidson, Sanjay Dubé, Larry Ereshefsky, Georges Gharabawi, Andrew C. Leon, Jean-Pierre Lepine, Steven G. Potkin, and An Vermeulen. What is causing the reduced drug-placebo difference in recent schizophrenia clinical trials and what can be done about it? *Schizophrenia bulletin*, 36(3):504–509, 2008.
- [17] TP Laughren. The scientific and ethical basis for placebo-controlled trials in depression and schizophrenia: an fda perspective. *European Psychiatry*, 16(7):418–423, 2001.
- [18] Sandra Lee, John R Walker, Laura Jakul, and Kathryn Sexton. Does elimination of placebo responders in a placebo run-in increase the treatment effect in randomized clinical trials? a meta-analytic evaluation. *Depression and anxiety*, 19(1):10–19, 2004.
- [19] Yuyin Liu, Denis Rybin, Timothy C Heeren, and Gheorghe Doros. Comparison of novel methods in two-way enriched clinical trial design. *Statistics in medicine*, 38(21):4112–4130, 2019.
- [20] Craig H Mallinckrodt, Adam L Meyers, Apurva Prakash, Douglas E Faries, and Michael J Detke. Simple options for improving signal detection in antidepressant clinical trials. *Psychopharmacology bulletin*, 40(2):101, 2007.
- [21] Craig H Mallinckrodt, Roy N Tamura, and Yoko Tanaka. Recent developments in improving signal detection and reducing placebo response in psychiatric clinical trials. *Journal of psychiatric research*, 45(9):1202–1207, 2011.
- [22] Bengt Muthén and Hendricks C Brown. Estimating drug effects in the presence of placebo response: causal inference using growth mixture modeling. *Statistics in medicine*, 28(27):3363–3385, 2009.

- [23] David Nutt and Guy Goodwin. Ecnp summit on the future of cns drug research in europe 2011: report prepared for ecnp by david nutt and guy goodwin. *European Neuropsychopharmacology*, 21(7):495–499, 2011.
- [24] Mary O’Hara and Pamela Duncan. Why ’big pharma’ stopped searching for the next prozac. <https://www.theguardian.com/society/2016/jan/27/prozac-next-psychiatric-wonder-drug-research-medicine-mental-illness>, 2016. Accessed: 2019-08-07.
- [25] Menelas N Pangalos and Christopher C Gallen. Drug discovery for disorders of the central nervous system. *Neurotherapeutics*, 2(4):539–540, 2005.
- [26] Pharmafile. The brain drain. <http://www.pharmafile.com/news/172099/brain-drain>, 2012. Accessed: 2019-08-07.
- [27] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.
- [28] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- [29] Denis Rybin, Gheorghe Doros, Michael J Pencina, and Maurizio Fava. Placebo non-response measure in sequential parallel comparison design studies. *Statistics in Medicine*, 34(15):2281–2293, 2015.
- [30] Denis Rybin, Robert Lew, Michael J Pencina, Maurizio Fava, and Gheorghe Doros. Placebo response as a latent characteristic: Application to analysis of sequential parallel comparison design studies. *Journal of the American Statistical Association*, 113(524):1411–1430, 2018.
- [31] Elizabeth Y Shang, Megan A Gibbs, Jaren W Landen, Michael Krams, Tanya Russell, Nicholas G Denman, and Diane R Mould. Evaluation of structural models to describe the effect of placebo upon the time course of major depressive disorder. *Journal of pharmacokinetics and pharmacodynamics*, 36(1):63–80, 2009.
- [32] Mark Sinyor, Anthony J Levitt, Amy H Cheung, Ayal Schaffer, Alex Kiss, Yekta Dowlati, and Krista L Lanctôt. Does inclusion of a placebo arm influence response to active antidepressant treatment in randomized controlled trials? results from pooled and meta-analyses. *Journal of Clinical Psychiatry*, 71(3):270–279, 2010.
- [33] Roy N Tamura and Xiaohong Huang. An examination of the efficiency of the sequential parallel design in psychiatric clinical trials. *Clinical Trials*, 4(4):309–317, 2007.
- [34] Roy N Tamura, Xiaohong Huang, and Dennis D Boos. Estimation of treatment effect for the sequential parallel design. *Statistics in medicine*, 30(30):3496–3506, 2011.
- [35] Madhukar H Trivedi and John Rush. Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications? *Neuropsychopharmacology*, 11(1):33, 1994.

- [36] B Timothy Walsh, Stuart N Seidman, Robyn Sysko, and Madelyn Gould. Placebo response in studies of major depression: variable, substantial, and growing. *Jama*, 287(14):1840–1847, 2002.
- [37] Xiangmin Zhang, Yeh-Fong Chen, and Roy Tamura. The plan of enrichment designs for dealing with high placebo response. *Pharmaceutical statistics*, 17(1):25–37, 2018.

Curriculum Vitae

