A STUDY OF NON-REGULARITY IN DYNAMIC TREATMENT REGIMES AND SOME DESIGN CONSIDERATIONS FOR MULTICOMPONENT INTERVENTIONS

by

Bibhas Chakraborty

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2009

Doctoral Committee:

Professor Susan A. Murphy, Chair
Professor Roderick J. Little
Professor Vijayan N. Nair
Professor Victor J. Strecher

To the loving and inspiring memory of my grandfather

ACKNOWLEDGEMENTS

I would like to express my heart-felt gratitude to my advisor Professor Susan Murphy for her help and guidance, and for pushing me to strive for excellence in research. I am deeply indebted to Professor Vijay Nair; he has been very kind and resourceful to me throughout my graduate school years at Michigan. I would like to thank my collaborators – Professor Victor Strecher, Professor Linda Collins, and Assistant Professor Suzanna Zick – who helped me develop my interdisciplinary research skills. My special thanks to Professor Roderick Little for kindly agreeing to be on my doctoral committee, and for helpful remarks about my research.

I acknowledge the support from the National Institute of Health grants RO1 MH080015 and P50 DA10075 (Professor Murphy), and P50 CA101451 (Professor Strecher). My special thanks to Carola Carlier, Janine Konkel and Mike Nowak – data management and behavioral science staff members from the Center for Health Communications Research. Staff member Rhonda Moats from the Institute for Social Research, and several staff members from the Department of Statistics – Amy Rundquist, Lu Ann Custer, Mary Ann King and Suleman Diwan – deserve special mentioning as well. I am grateful to all my teachers and my friends, who are simply too many to list here.

Last but not the least, I want to thank my family – my parents, my sister, my grandmother and my late grandfather – for their support and encouragement in pursuing my doctoral studies.

TABLE OF CONTENTS

DEDICATION . . . . . **ii**

ACKNOWLEDGEMENTS . . . . . **iii**

LIST OF FIGURES . . . . . **vi**

LIST OF TABLES . . . . . **vii**

CHAPTER

I. Introduction 1

1.1 Designing Multicomponent Intervention Trials 1
1.1.1 Multiphase Optimization Strategy (MOST) 3
1.1.2 MOST and Factorial Designs 4
1.2 Dynamic Treatment Regimes 6
1.2.1 The Framework of Multi-stage Studies 7
1.2.2 Estimation and Inference 9
1.2.3 Dynamic Treatment Regimes and Multicomponent Interventions . 11
1.3 Outline 11

II. The Motivating Study on Smoking Cessation 14

2.1 Project Quit Study 14
2.1.1 Factors and Study Design 15
2.1.2 Data Collection 16
2.1.3 Outcome Measures 17
2.2 Forever Free Study 18
2.2.1 Factors and Study Design 19
2.2.2 Data Collection 20
2.2.3 Outcome Measures 21
2.3 Dynamic Treatment Regime Framework 22
2.3.1 Preliminary Descriptive Numbers 23

III. Developing Multicomponent Interventions using Fractional Factorial Designs 25

3.1 Introduction 25
3.2 Factorial Designs for Screening Studies 27
3.2.1 Addressing Criticisms against Factorial Designs 29
3.3 Operationalization of Screening Trials 46
3.4 Follow-up Studies 52
3.5 Discussion 56
3.6 Appendix A: Defining Relation and Resolution of an FFD 58

3.7 Appendix B: Relative Efficiency of The Two Ways of Forming The Test Statistic . . . 59

IV. Comparison of the MOST Approach and a Single Randomized Trial for Developing Multicomponent Interventions . . . 61

4.1 Introduction . . . 61
4.2 Methods . . . 62
4.2.1 Overview of the Simulation . . . 62
4.2.2 Data Generation Model . . . 63
4.2.3 Experimental conditions . . . 66
4.2.4 Operationalization of the classical and the MOST approaches . . . 67
4.3 Results . . . 70
4.4 Discussion . . . 74
4.4.1 Limitations . . . 80
4.4.2 Conclusions . . . 80
4.5 Appendix A: Data Generating Model . . . 81
4.6 Appendix B: Operationalization of the MOST Approach . . . 85
4.7 Appendix C: Summary Results across Different Simulation Conditions . . . 93

V. Inference for Non-regular Parameters in Optimal Dynamic Treatment Regimes . . . 96

5.1 Introduction . . . 96
5.2 Estimation and Inference via Q-learning . . . 99
5.2.1 Notation and Data Structure . . . 99
5.2.2 Q-learning with Linear Models . . . 100
5.2.3 The Inference Problem . . . 101
5.2.4 Non-regularity in Inference . . . 102
5.3 Different Regularized Estimators . . . 104
5.3.1 Hard-threshold Estimator . . . 104
5.3.2 Soft-threshold or Shrinkage Estimator . . . 106
5.4 Simulation Study . . . 108
5.4.1 Results . . . 114
5.5 Analysis of the Smoking Cessation Data . . . 119
5.5.1 Complete-case Analysis . . . 120
5.5.2 Analysis using Multiple Imputation . . . 123
5.6 Discussion . . . 132
5.7 Appendix A: Proof of Lemma V.1 . . . 136
5.8 Appendix B: Proof of Lemma V.2 . . . 140
5.9 Appendix C: A Very Brief Review of Multiple Imputation . . . 142

VI. Future Work and Conclusion . . . 147

6.1 Follow up Studies for Multicomponent Interventions . . . 147
6.2 Soft-threshold Estimator for More than Two Treatments per Stage . . . 148
6.3 Consistent Bootstrap Procedure for Non-regular Settings . . . 149
6.4 Concluding Remarks . . . 149

BIBLIOGRAPHY . . . **151**

LIST OF FIGURES

Figure

4.1 Data generation model for simulation. 64

5.1 Hard-threshold and Soft-threshold pseudo-outcomes compared with the Hard-max pseudo-outcome. 107

5.2 Interaction plots: (a) source by self-efficacy (left panel), (b) story by education (right panel), along with confidence intervals for predicted stage 1 pseudo-outcome. 123

5.3 Histograms of PQ6Quitstatus, PQ6MonthsNS, and PQ6NumOfAttempts. 125

5.4 Histograms of PQ6NumOfAttempts and its square-root. 126

5.5 Histograms of FF6Quitstatus, FF6MonthsNS, and FF6NumOfAttempts. 127

5.6 Histograms of FF6NumOfAttempts and its square-root. 128

5.7 Histograms of PQ6MonthsNS and FF6MonthsNS (trinary). 128

5.8 Auto-correlation function of the worst cosine function. 129

5.9 Auto-correlation functions of the regression parameters associated with the stage-1 outcome (PQ6MonthsNS). 130

5.10 Auto-correlation functions of the regression parameters associated with the first dummy variable corresponding to the stage-2 outcome (FF6MonthsNSDummy1). . 130

5.11 Auto-correlation functions of the regression parameters associated with the second dummy variable corresponding to the stage-2 outcome (FF6MonthsNSDummy2). . 131

LIST OF TABLES

Table

2.1 Project Quit Factors and Their Levels 15

2.2 Two Forever Free Factors and Their Levels 19

2.3 Descriptive Numbers about the Smoking Cessation Data 23

3.1 Power to screen A_1 in absence and presence of an interaction 44

3.2 Recommended resolution IV FFDs under varying anticipated interactions 49

4.1 Mean Intervention Outcome under Classical and MOST Approaches (averaged over 1000 simulated data sets) 71

4.2 Comparison of Classical and MOST Approaches on E(Y) (Percentage of Data Sets) 71

4.3 Accuracy of Component Selection under Classical and MOST Approaches (Percentage of Data Sets) 72

4.4 Whether the Classical (C) or the MOST (M) approach produced the largest value of $E(Y)$ under a variety of simulation conditions (This is a summary across 9 simulation settings of the entries that correspond to the Table 1 in the main text). 94

4.5 Whether the Classical (C) or the MOST (M) approach produced a higher $E(Y)$ value in more data sets than its counterpart under a variety of simulation conditions (This is a summary across 9 simulation settings of the entries that correspond to the Table 2 in the main text). 94

4.6 Whether the Classical (C) or the MOST (M) approach showed more accuracy in component selection under a variety of simulation conditions (This is a summary across 9 simulation settings of the entries that correspond to the Table 3 in the main text). 95

5.1 Distribution of the linear combination $(\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1)$ 110

5.2 Summary statistics and coverage rates of 95% and 90% nominal percentile (PB), hybrid (HB), and double (DB) bootstrap CIs for ψ_{10} using the hard-max (HM), the hard-threshold with $\alpha = 0.08$ ($\text{HT}_{0.08}$) and $\alpha = 0.2$ ($\text{HT}_{0.20}$), and the soft-threshold (ST) estimators. A "*" indicates significantly different coverage rate than the nominal rate, using a test of proportion (Type I error rate = 0.05). 116

5.3 Summary statistics and coverage rates of 95% and 90% nominal percentile (PB), hybrid (HB), and double (DB) bootstrap CIs for ψ_{10} using hard-max (HM), hard-threshold with $\alpha = 0.08$ ($HT_{0.08}$) and $\alpha = 0.2$ ($HT_{0.20}$), and soft-threshold (ST) estimators. A "*" indicates significantly different coverage rate than the nominal rate, using a test of proportion (Type I error rate = 0.05). 118

5.4 Regression coefficients and 95% hybrid bootstrap confidence intervals at stage 1, using both the hard-max and the soft-threshold estimators. 122

5.5 Baseline Variables subject to Missingness 124

5.6 Stage-1 (collected at 6 months) Variables subject to Missingness 124

5.7 Stage-2 (collected at 12 months) Variables subject to Missingness 126

5.8 Regression coefficients and 95% hybrid bootstrap confidence intervals at stage 1, using both the hard-max and the soft-threshold estimators, and using multiple imputation. 132

5.9 Length of Bootstrap CIs. 132

CHAPTER I

Introduction

In this dissertation, we investigate two problems: (1) the problem of establishing a "gold standard" for developing and optimizing multicomponent interventions that allows for valid inference about individual components and their interactions (Ch. III, IV); and (2) the problem of *non-regularity* that arises in the estimation of optimal *dynamic treatment regimes* from longitudinal data on patients (Ch. V). Both these methodological research directions are motivated by our involvement in the design and analysis of a smoking cessation trial [88] conducted by the Center for Health Communications Research (CHCR) at the University of Michigan; a description of this trial is given in Chapter II.

1.1 Designing Multicomponent Intervention Trials

The first problem investigated in this dissertation is the design of multicomponent intervention trials. Multicomponent or complex interventions are increasingly being used in many health domains, e.g. AIDS [41], cardiovascular diseases [28], depression [105], diabetes [65], drug abuse [68], gerontology [1], obesity [8], and smoking cessation [88]. While some components may involve a medication, many components are behavioral, implementation, or delivery factors [28, 105, 88]. As has been rec-

ognized in the literature [1, 18, 39, 86, 103], development and evaluation of these multicomponent interventions pose additional design challenges over those of single-component interventions, and these challenges tend to be addressed poorly by the standard two-group randomized controlled trials. More specifically, such two-group randomized trials do not provide direct information on which components are active, whether they have been set at optimal levels, and whether any of the components interact.

The classical approach and its problem

The classical approach (sometimes called the *treatment package strategy*) to develop multicomponent interventions [47, 103, 102] consists of constructing a likely best intervention package based primarily on prior empirical research, readings of the literature, theory and clinical experience. This intervention is then evaluated in a standard randomized controlled trial. In the course of this trial, data are collected not only on the outcomes of primary interest but also on other variables so as to enable quasi-experimental, non-experimental and *post hoc* analyses [68, 1, 96, 97] aimed at shedding light on what worked well and what might need improvement. Examples of such analyses include regressing outcomes on naturally occurring variation in participation, compliance, or implementation fidelity. Other observational analyses include theory-based mediational analyses [103, 106]. The intervention is often refined based on the findings of these analyses, and then the refined version is tested in another two-group randomized trial. Sometimes several such iterations are performed to refine the multicomponent intervention. Thus the questions regarding individual components and their interactions are answered by observational analyses.

The main problem with this approach is that it depends heavily on the non-

experimental, observational analyses. As is well-known [44, 72, 74], findings that are not based on randomization are contaminated by the likely presence of unknown *confounders* (e.g. the variables that affect both the receipt of an intervention component and the outcome). As a consequence, the effects of individual components and interactions may be misinterpreted resulting in a suboptimal intervention.

1.1.1 Multiphase Optimization Strategy (MOST)

To address the above problem with the classical approach, an experimental approach called the *Multiphase Optimization Strategy* (MOST) was proposed by Collins et al. [25], and further developed and illustrated by Nair et al. [62]. Examples of successful implementation of MOST include the works of Nair et al. [62] in breast cancer prevention and Strecher et al. [88] in smoking cessation. A description of this smoking cessation trial can be found in Chapter II.

The MOST approach includes additional evidentiary steps along with the two-group randomized trial as part of the process of building and evaluating multicomponent interventions. MOST consists of two evidentiary phases to precede and inform a *confirmatory* two-group randomized trial. The first phase, called *screening*, consists of randomized experimentation designed to obtain estimates of the effects of individual components and selected interactions between components. The resulting experimental evidence provides the basis for preliminary decisions about which components to select for inclusion. A second phase of additional experimentation, called *refining* (sometimes called *follow-up* study), is used to identify the best level of one or more components, to investigate interactions between components, and to resolve any other remaining questions. The final confirmatory two-group trial is sometimes referred to as the *confirming* phase of the MOST approach. Information on cost and

burden can be collected in the course of experimentation and included when decisions are made concerning choices of components and/or levels. Conclusions drawn from the results of the screening and refining phases form the basis for specification of an intervention (to be tested in the confirmatory trial) that consists of a set of active components implemented at levels selected to maximize efficacy, effectiveness, and/or cost-effectiveness.

Even though the screening and refining studies precede a confirmatory randomized trial of the "optimized" multicomponent intervention vs. control, they are not pilot studies by most widely accepted definitions [76, 48, 98]. According to these definitions, a pilot study is typically conducted to assess the feasibility (of recruitment, intervention delivery, data collection) of a full-blown randomized trial; indeed pilot studies may be conducted with little regard for power, and may not even involve randomization. By contrast, screening and refining studies as described by Collins et al. [25] are adequately powered randomized trials intended to assist in refining and optimizing multicomponent interventions and may themselves be preceded by pilot studies to assess feasibility.

1.1.2 MOST and Factorial Designs

Full and fractional factorial designs (FFD) that have been extensively used in agricultural and industrial experiments for many decades [108, 36, 9, 107] can be used efficiently in the screening phase of MOST. Nair et al. [62] described the use of FFDs within the MOST framework in two behavioral intervention studies, e.g. *Guide to Decide* and *Project Quit*. One goal of the current dissertation is to establish FFDs within the MOST framework as the "gold standard" for developing multicomponent interventions. Below we describe how we will proceed to achieve this goal.

Justifying the use of FFDs in screening studies

Even though the use of FFDs in the screening phase of MOST was described by earlier authors [25, 62, 88], there exists a tremendous amount of controversy regarding the use of factorial designs in the clinical and behavioral intervention trials literature [16, 103, 67]. Factorial designs have been assessed by biostatisticians in the context of confirmatory medication trials; the objective has been to *evaluate* the usefulness of a combined medication over a single medication. In this context, the criticisms of factorial designs relate to cost, feasibility, ethics, toxicity of combined medications, interpretation of main effects in presence of active interactions, and power to detect interactions. In Chapter III of the current dissertation, which is based on a recent paper [19], we describe how one can address these common criticisms (and some misconceptions) regarding the use of (fractional) factorial designs in the context of screening trials (which is different from evaluation or confirmatory trials) for developing multicomponent interventions. To the best of our knowledge, addressing the criticisms against FFDs in the context of the MOST framework (screening phase) has not been done before.

Additionally in Chapter III, we provide an operationalization of screening studies using FFDs for up to six components. We also present some hypothetical examples of follow-up studies, as well as the follow-up study design of *Project Quit*.

Comparing MOST with the classical approach in a simulation study

MOST is a relatively new experimental framework to a lot of behavioral intervention scientists. We feel that an illustrative simulation study, which shows that MOST performs better than the classical approach under a lot of simulated scenarios

that mimic some of the well-known characteristics of real studies, has the promise of disseminating this framework further. In Chapter IV, which is based on a recent paper [22], we present such an illustrative simulation study that provides a head-to-head comparison of MOST (implemented using FFDs) with the classical approach (two-group randomized trial followed by observational analyses). This comparison is based on a generative model that involves five intervention components, varying levels of adherence to each component, a negative interaction between two components, an unknown confounder, and a continuous outcome variable. The simulation results show that under a lot of different scenarios (e.g. varying effect size and sample size), the MOST approach (implemented using FFDs) outperform the classical approach in terms of various criteria, e.g. optimizing the mean outcome of the final intervention and identifying the best intervention. This chapter, we believe, will strengthen the case for FFDs within the MOST framework as the "gold standard" for developing multicomponent interventions.

1.2 Dynamic Treatment Regimes

The second problem studied in this dissertation is the phenomenon of *non-regularity* arising in the estimation of the optimal *dynamic treatment regimes* (DTR). DTRs are useful tools in treating chronic disorders (e.g., depression, schizophrenia, substance abuse, HIV infection etc.). In order to manage the waxing and waning nature of these disorders, clinicians typically treat patients in multiple stages, adapting the treatment type and dosage to the ongoing measures of an individual patient's response, burden, adherence to prior treatment, side effects, and preference. Practice guidelines for clinicians offer treatment recommendations, but they often rely

heavily on expert opinion because of limited direct scientific evidence. DTRs represent a paradigm to operationalize this adaptive clinical practice, and thereby to improve it. A DTR is a sequence of (treatment) decision rules, one per stage. Each decision rule takes a patient's treatment and covariate history as input, and outputs a recommended treatment. To make these treatment rules "evidence-based", one has to estimate them from longitudinal data on patients in a principled way. Data for estimating DTRs come from multi-stage studies – either an observational longitudinal study or a *sequential multiple assignment randomized trial* (SMART) [49, 50, 32, 58, 59]. In SMART designs, each patient is followed through stages of treatment and at each stage the patient is randomized to one of the possible treatment options. Experimental designs similar to SMART have been implemented in the treatments of schizophrenia [81], depression [78] and cancer [93, 99]. Below we briefly discuss the framework of multi-stage studies.

1.2.1 The Framework of Multi-stage Studies

For simplicity, let us focus on studies with only two stages of treatment. Longitudinal data on a single patient are given by the trajectory

$$\{O_1, A_1, O_2, A_2, O_3\},$$

where O_j $(j = 1, 2)$ denotes the covariates observed prior to treatment at the beginning of the j-th stage, O_3 is the observation at the end of stage 2, and A_j $(j = 1, 2)$ is the "action" or treatment assigned at the j-th stage subsequent to observing O_j. Define history available at each stage as: $H_1 = O_1$, $H_2 = (O_1, A_1, O_2)$. The study can have either a single primary outcome Y observed at the end of stage 2, or two outcomes Y_1, Y_2 observed at the two stages (and the interest is in $Y_1 + Y_2$). Note that

the case of a single outcome Y observed at the end can be viewed as a case with $Y_1 \equiv 0$ and $Y_2 = Y$. We assume that the outcomes are specified summaries of observations and treatments, i.e., $Y_1 = f_1(O_1, A_1, O_2)$ and $Y_2 = f_2(O_1, A_1, O_2, A_2, O_3)$, with known functions f_1, f_2. In this simple framework, a DTR is a two-component vector of decision rules, say (d_1, d_2), with $d_j(H_j) \in \mathcal{A}_j$, where $\mathcal{A}_j$ represents the set of possible treatments at the j-th stage. An example of such a decision rule can be: "stop treatment if $\psi^T H_j > 0$, otherwise maintain on current treatment", where ψ is a vector of parameters. A DTR is called optimal if it leads to a maximal mean Y (or, maximal mean sum of Y_j's). The goal is to construct optimal DTRs, e.g. by estimating the ψ's featuring in the optimal DTR. Note that this simple framework of two stages can be generalized to more than two stages. Below we present two concrete examples.

Two-stage Smoking Cessation Trial (Strecher et al., 2008)

We present a randomized, two-stage, longitudinal, internet-based smoking cessation and relapse prevention study. We will treat this example in great detail in Chapters II and V; here is only a gentle overview. The stage 1 of this study (*Project Quit*) was conducted to find an optimal multicomponent behavioral intervention to help adult smokers quit smoking; and the stage 2 (*Forever Free*) was a follow-on study to help those (among the participants of *Project Quit*) who already quit stay quit, and help those who failed at the previous stage with a second chance. Here O_1 consists of baseline variables (e.g., motivation, self-efficacy, education), O_2 (also Y_1) consists of several stage 1 outcomes (e.g., quit status, reduction in the average number of cigarettes smoked per day, number of months not smoked during the study period – all measured at 6 months from the baseline), and O_3 (also Y_2) consists

of the same outcome variables measured at stage 2 (12 months from the baseline). A_1 and A_2 represent the behavioral interventions given at stages 1 and 2 respectively.

Two-stage Cancer Trials (Wahed and Tsiatis, 2004)

In a two-stage design for cancer trials, patients are often treated initially with an induction therapy (powerful chemotherapy to induce remission of disease) followed by (at some later time-point) either some maintenance therapy to intensify or augment the effects of induction therapy, if the patient "responds" (shows sign of remission), or some other induction therapy (switch of treatment), if the patient does not respond to the initial therapy. Here O_1 stands for pre-treatment variables (e.g., age, sex, ethnicity etc.), A_1 is the induction therapy at the first stage, O_2 is whether or not the patient "responds" to initial induction therapy, A_2 is either the maintenance therapy (if the patient responds) or the new induction therapy at the second stage (if the patient does not respond), and O_3 (also the outcome Y) is the disease-free survival time.

1.2.2 Estimation and Inference

Methodological developments for estimating DTRs took place in recent times. Thall et al. [93, 94, 95] considered likelihood-based methods (both frequentist and Bayesian) for estimating DTRs, primarily in the context of cancer. Murphy et al. [60] provided a method of estimation for the mean outcome that would have been observed had the study population followed a particular DTR, based on observational longitudinal data and under the assumption of *sequential randomization*. Parmigiani [64] considered modeling medical decisions by Bayesian approaches. Murphy [56] provided a semi-parametric regret modeling methodology, sometimes called A-

Learning [7]. An efficient version of this method was provided by Robins [71]. A good discussion of the relationship between these methods can be found in Moodie et al. [55]. Other methods for estimating DTRs include the semi-parametric methods due to Lunceford et al. [51] and Wahed and Tsiatis [99, 100] in the context of two-stage cancer trials, with survival distribution as the primary outcome. Rosthoj et al. [73] considered a case study based on Murphy's [56] methodology using observational data. The problem of inference for the parameters of the optimal DTR was studied by [71].

Non-regularity

Robins [71] identified the problem of *non-regularity* in the context of estimating optimal DTRs. As discussed by Robins, the treatment effect parameters at any stage prior to the last can be *non-regular* under certain longitudinal distributions of the data which he called *exceptional laws*. By non-regularity, we mean that the asymptotic distribution of the estimator of the treatment effect parameter does not converge uniformly over the parameter space. This phenomenon of non-regularity causes bias in estimation, and leads to poor frequentist properties (e.g. coverage rates) of the confidence intervals. Recently Moodie and Richardson [54] provided a method called *Zeroing Instead of Plugging In* (ZIPI) for correcting the bias in the estimation of the optimal DTRs resulting under exceptional laws.

In Chapter V of this dissertation, we illustrate the problem of non-regularity using a method of estimation called Q-learning [101, 90, 58]. This method, like Robins' *g-estimation of optimal structural nested mean models*, suffers from non-regularity – the common reason being an underlying non-smooth maximization operation. We show that under simple conditions, Q-learning is equivalent to an inefficient version

of Robins' method; however Q-learning is simpler to visualize. To address non-regularity, we propose a new estimator called the *soft-threshold* estimator within the framework of Q-learning, and compare it with other available estimation techniques that attempt to address this problem (including the ZIPI or *hard-threshold* estimator) via extensive simulations. The soft-threshold estimator falls within the class of *shrinkage* estimators, and the tuning parameter governing the shrinkage is specified by an *empirical Bayes* formulation of the problem. The content of this chapter is based on a recent paper by Chakraborty et al. [20].

1.2.3 Dynamic Treatment Regimes and Multicomponent Interventions

Note that while developing a DTR, one has to decide on a number of questions, e.g., when to start treatment, which treatment type and/or dosage to start with, when to step up (augment) treatment and to which, when to step down treatment to maintenance or monitoring therapy, and what information to use to make the above decisions. Thus a DTR can be viewed as a possibly high-dimensional multicomponent intervention. Hence a series of developmental or screening trials may be necessary before conducting a confirmatory trial to evaluate a DTR in comparison with a standard alternative treatment (control). Hence the MOST framework in general, and FFDs in particular, can provide useful tools in the development of DTRs. This conceptual connection, in a way, ties the two broad problems considered in this dissertation together.

1.3 Outline

In the present chapter, we have discussed the MOST framework for developing multicomponent interventions. We have also introduced the general concept of DTRs and the phenomenon of non-regularity that arises in the estimation of DTRs. In the subsequent chapters, we will build on this background and discuss specific research done along these lines. In Chapter II, we give a detailed description of a study on smoking cessation [88]. We use the first stage of this study (*Project Quit*) as an example where MOST (using FFDs) was successfully implemented. Also in this smoking cessation study, the problem of estimating the optimal DTR arises naturally. We use this as a motivating example for the methodology considered in Chapter V.

Chapter III is devoted primarily to addressing the concerns and criticisms against FFDs in the context of the screening phase of MOST. Here we also provide an operationalization of screening trials using FFDs and some discussion on follow-up studies. Chapter IV provides a head-to-head comparison between two competing approaches to develop multicomponent interventions, e.g., (a) MOST using FFDs, and (b) traditional two-group randomized trial followed by non-experimental post hoc analysis, using a complex simulation study. The simulation study shows how approach (a) outperforms approach (b) in most scenarios.

In chapter V, we present the problem of non-regularity arising in the estimation of optimal DTRs in full detail. Here we derive Q-learning as an inefficient version of Robins' [71] method of estimation in *structural nested mean models* (SNMM). We propose the *soft-threshold* estimator to address non-regularity, and derive it as the empirical Bayes estimator under a hierarchical Bayesian model. We also present a comprehensive simulation study to compare different methods in regular as well as

non-regular settings. Finally, we provide two separate analyses (one is the *complete-case analysis*, the other using *multiple imputation*) of the smoking cessation data to illustrate the developed methodology.

In chapter VI, we discuss possible extensions and future work. With reference to the work on designing multicomponent intervention trials, future research consists of exploring the area of follow-up studies with more rigor. Future research directions related to DTRs include extending the proposed soft-threshold method to the setting of more than two treatment options per stage, and devising a consistent bootstrap procedure for using in non-regular settings. We briefly outline our plans to explore these in future.

CHAPTER II

The Motivating Study on Smoking Cessation

In this chapter, we will describe the *Project Quit* and the *Forever Free* studies, conducted by the Center for Health Communications Research (CHCR) at the University of Michigan. Together they provide a setting where the problem of estimating optimal dynamic treatment regime arises naturally; we want to use this setting as an example to motivate the methods proposed later in this thesis. The Project Quit study also exemplifies a successful implementation of the MOST framework and the use of fractional factorial designs (FFDs) therein.

2.1 Project Quit Study

Project Quit is a web-based smoking cessation program developed by researchers at the Center for Health Communications Research of the University of Michigan, Ann Arbor, and two health maintenance organizations (HMO), e.g. the Group Health Cooperative (GHC), Seattle, and the Henry Ford Health System (HFHS), Detroit. The funding for this study was provided by the National Cancer Institute (NCI). The subjects of this study were recruited from the population of adult patients of two health maintenance organizations (HMOs), e.g. GHC and HFHS, associated with NCI's Cancer Research Network (CRN). This population provided

a broad representation of ethnicity, gender, age, health status, and geography. The collaboration allowed the researchers to test some cutting-edge web-based technology in a real-world environment that has the infrastructure for both evaluating and disseminating population-based cancer prevention and control programs. This study used the MOST framework (implemented using FFDs) to screen and identify the effects of some psycho-social and communication factors (treatment components) influencing smoking cessation.

2.1.1 Factors and Study Design

An FFD was used to screen and identify the factors influencing cessation from a set of potentially active factors of a web-based cessation induction intervention. Six factors listed below, each at two levels, were considered originally.

Table 2.1: Project Quit Factors and Their Levels

Factor	High Level (+)	Low Level (−)
Outcome Expectations	High Tailoring Depth	Low Tailoring Depth
Efficacy Expectations	High Tailoring Depth	Low Tailoring Depth
Success Stories	High Tailoring Depth	Low Tailoring Depth
Message Source	High Personalization	Low Personalization
Message Framing	Gain Framing	Loss Framing
Message Exposure	Single	Multiple

The originally planned design was a 32-arm FFD. However, due to an implementation error, a decision was made later to drop the factor *Message Framing* from the study, and subsequently fold the design to have 16 treatment arms. Details of the study design can be found in [88] and also in chapter III.

The factors of this study were examined across a set of individual characteristics. These characteristics were baseline level of motivation and self-efficacy (on a 10-point

scale), barriers (e.g., fear of weight gain, smoking habit of spouse, prevalence of friends and co-workers who smoke etc.), need for cognition (an individual's tendency to engage in and enjoy effortful cognitive endeavors), health locus of control (refers to the belief that behaviors are causally related to outcomes, or are determined by external factors such as luck, chance etc.), and socio-demographic (e.g., age, gender, education, ethnicity) and health status characteristics (e.g., hypertension, diabetes, heart disease etc.). Interactions of some of these subject characteristics with the factors were studied.

2.1.2 Data Collection

During the subject recruitment phase, an introductory recruitment letter was sent to individuals randomly chosen from the study population. The letter explained the project (describing the research, its purpose, and its risks and benefits) and asked individuals to voluntarily participate in this project by visiting the web site

`http://www.projectquit.org`

to complete an electronic consent form and eligibility questionnaire. Subjects were eligible for the study if they satisfied the following criteria:

1. They were between 21 and 70 years of age.
2. They were assigned to a physician panel at GHC or HFHS as of January 1, 2003.
3. They were capable of communicating in English.
4. They were current smokers, defined as: they had smoked at least 100 cigarettes in their lifetime, they reported smoking in the last seven days, and they reported

smoking at least 10 cigarettes per day during the time of this survey.

5. They were not enrolled in any other smoking cessation program, or not using pharmacological therapy for smoking cessation.

6. They had access to the internet and had an email account that they used to use at least twice a week.

All eligible individuals who agreed to participate in the study completed a baseline survey, selected a quit date between two and four weeks from the date of the baseline survey completion date, and were randomized to one of the treatment arms.

The baseline questionnaire assessed the subject's smoking history, psychosocial, health, and demographic characteristics relevant to smoking cessation. Randomization was done automatically by the computer and was invisible to the subject. Data from the baseline survey were used to generate the experimental web site condition to which the subject was randomly assigned. Subjects were encouraged to complete the baseline questionnaire through the use of periodic email reminders. Follow-up assessment was performed six months after initial login. At this time, the subject was called by a trained telephone interviewer using a computer-assisted telephone interview (CATI).

2.1.3 Outcome Measures

The primary outcome in the Project Quit study was the seven-day point-prevalence in smoking cessation (binary) at six months following the baseline assessment. During the 6-month evaluation survey, each subject was asked if s/he had smoked any cigarettes, even a puff, in the last seven days. A subject answering "yes" to this

question was marked as a smoker.

A criticism of the above outcome measure is that it is only based on a subject's smoking status in last seven days, as opposed to last 6 months. However, data on some secondary outcomes such as number of months not smoked by a subject, changes in the number of cigarettes smoked per day and number of quit attempts in the past 6 months were also collected.

2.2 Forever Free Study

Forever Free is a web-based smoking cessation and relapse prevention program developed as a follow-on program to the Project Quit study. The funding for this developmental project was provided by the University of Michigan. The goal of this study was to help the participants of the Project Quit study – to help the quitters stay quit and to help the non-quitters continue the quitting process. Subjects of this study were recruited from the participants of the Project Quit study. During the six-month CATI follow-up to Project Quit treatment assignment, the participants were asked if they wanted to participate in the Forever Free program and if they would mind being contacted about it. The subjects who consented on the phone, received invitation email for participating in Forever Free. Then the subjects gave their consents online by visiting the Forever Free web site. The subjects who consented were subsequently randomized to one of the 5 treatment arms discussed below. Subjects were blocked by their quit status (based on their seven-day point prevalence in the Project Quit study, HMO membership (GHC vs. HFHS), and exposure in Project Quit (single or multiple).

2.2.1 Factors and Study Design

In the Forever Free study, two factors, each at two levels, gave rise to a 2×2 full factorial design. Additionally, a control arm was considered for certain comparisons of interest. Thus a total of 5 treatment arms were constructed. If n be the total number of participants in this study, then the planned sample size for the control arm was $\frac{n}{3}$, while each of the other 4 arms were planned to have a sample size of $\frac{n}{6}$.

Table 2.2: Two Forever Free Factors and Their Levels

Factor	High Level (+)	Low Level (−)
Tailoring Locus	Expert System	User Navigation
Graphic Content	Graphic Rich	Graphic Poor

The intervention primarily consisted of the contents of 8 smoking relapse prevention booklets, written by Dr. Thomas Brandon, and available on the NCI web site:

`http://www.smokefree.gov/pdf.html`

The wordings of these booklets were adjusted to reflect that a participant had or had not quit (according to Project Quit outcome). Each of these booklets is about handling a particular situation when a subject may feel the urge to smoke, e.g., stress, loneliness etc. The factors determined how these booklets were presented to the subjects. For example, with reference to the second factor, the graphic poor version (lower level) included simple clip-art style graphics found in the PDF version of the booklets, while the graphic rich version (higher level) included enhanced graphics that better communicated the ideas in the text. Thus there were four versions of each booklet, viz., graphic rich for those who have quit smoking, graphic poor for

those who have quit smoking, graphic rich for those who have not quit smoking, and graphic poor for those who have not quit smoking. On the other hand, the first factor, tailoring locus, determined whether or not the choice of booklets were tailored to a particular subject's needs. In lower level of this factor (user navigation), the subjects were presented with an web page with links to all the 8 booklets, and they had to self-select which ones they wanted to review by clicking on appropriate links. In higher level of this factor (expert system), although the subjects had access to all the 8 booklets, they were advised to review up to 3 booklets by the "expert", based on their answers to a previous questionnaire. Subjects in the control arm did not get any of these booklets; they only received an encouraging message to quit smoking.

2.2.2 Data Collection

Subjects were invited to participate in this follow-on study as part of the Project Quit six-month CATI follow-up. They were sent an email reminder with a link to the site

`http://www.projectquit.org/foreverfree`

approximately seven days after completing the CATI interview. Information collected in the six-month CATI follow-up were used to tailor the selection of the appropriate booklet for participants in the expert system cells. If a participant failed to log in to the web site, they were sent email reminders at 1, 3, and 5 days reminding them to visit the web site to enroll in the follow-on study.

All subjects were sent an email reminder at three months (post enrollment in the follow-on study) asking them to return to the web site and complete a short survey.

If a participant failed to visit the web site to complete the survey, they were sent email reminders at 1, 3, and 5 days asking them to visit the web site to do the three-month survey. If a subject started the three-month survey but failed to complete it, they were sent email reminders at 1, 3, and 5 days after they started the three-month survey asking them to return and complete the three-month survey.

All subjects were sent an email reminder at six months (post enrollment in the follow-on study) asking them to return to the web site and complete the six-month survey. If a subject failed to visit the web site to complete the survey, they were sent email reminders at 1, 3, and 5 days asking them to visit the web site to do the six-month survey. If a participant started the six-month survey but failed to complete it, they were sent email reminders at 1, 3, and 5 days after they started the six-month survey asking them to return to the web site and complete the six-month survey. Subjects that did not complete the six-month survey on the web were mailed a copy of the survey to complete. After six months (post enrollment in the follow-on study), subjects were redirected to the NCI web site to view the booklets. Although both three-month and six-month data were collected, we will focus on the six-month data only.

2.2.3 Outcome Measures

The primary outcome in the Forever Free study was seven-day point-prevalence in smoking cessation, which is binary. At 3-month or 6-month survey, each subject was asked if s/he had smoked even a single cigarette during the last seven days. A subject answering "yes" to this question was marked as a smoker.

As before, secondary outcomes included how many months a subject has been off cigarettes, changes in the number of cigarettes smoked per day, and number of quit attempts.

2.3 Dynamic Treatment Regime Framework

Having described the study on smoking cessation, let us now identify the set-up for estimating a dynamic treatment regime. Note that, in the combined Project Quit – Forever Free study, two randomizations were done, one is the Project Quit treatment assignment (1 out of 16), the other is the Forever Free treatment assignment (1 out of 5). So these are the two time-points in the dynamic treatment regime framework. Ideally, for each subject, we have a longitudinal record of the form

$$\{O_1, A_1, O_2, A_2, O_3\},$$

where O_1 represents baseline subject characteristics, A_1 is the Project Quit treatment arm the subject was assigned to, O_2 stands for the subject characteristics and outcomes measured during 6-month CATI follow-up, A_2 is the Forever Free treatment arm the subject was assigned to, and O_3 stands for the subject characteristics and outcomes measured during 6-month follow-up from Forever Free treatment assignment. History at each stage can be defined as a suitable lower dimensional summary of previous O's and A's. The primary outcome at each stage is the (binary) seven-day point prevalence in smoking cessation. Some secondary outcomes (e.g., how long a subject has been off cigarettes) are also available.

The goal in the present problem is to estimate the optimal dynamic treatment regime (sequence of individualized treatments) that is most effective in smoking ces-

sation – the one that maximizes the combined outcome over the two stages, and also to attach measures of confidence (statistical inference).

For all subsequent analyses in this thesis, we will collapse the 4 different versions of the Forever Free intervention (due to little difference between them) to form a single treatment arm and compare it with the control arm.

2.3.1 Preliminary Descriptive Numbers

Here we present some preliminary descriptive numbers relating to the data. As we see below, a lot of subjects from Project Quit decided not to continue to Forever Free (only 479 out of 1848 subjects decided to continue); this step was part of the protocol, and hence these subjects are not considered as drop-out. However note that only 1401 out of 1848 stage-1 subjects completed the six-month CATI survey; these 1401 subjects are treated as *complete cases* and would be considered in *complete-case analysis* in chapter V; the remaining 447 subjects are considered drop-outs in the stage-1 data. Similarly, 281 subjects who completed the stage-2 six-month survey are considered as *complete cases* in the *complete-case analysis* in chapter V; the remaining $479 - 281 = 198$ are considered drop-outs in the stage-2 data. Also note that these numbers will change as we vary the outcome under consideration; here we focus on the primary outcome (quit status) only.

Table 2.3: Descriptive Numbers about the Smoking Cessation Data

Subjects at different stages of the study	n
Subjects who were assigned to a Project Quit treatment arm	1848
Subjects for whom Project Quit 6-month primary outcome is available	1401
Subjects who were assigned to a Forever Free treatment arm	479
Subjects for whom Forever Free 6-month primary outcome is available	281

Descriptive checks revealed that drop-out was more or less uniform across the different treatment arms in both stages. In chapter V, we will also present a refined analysis using *multiple imputation* (as is well-known, complete case analysis often gives biased results).

CHAPTER III

Developing Multicomponent Interventions using Fractional Factorial Designs

Multicomponent interventions composed of behavioral, delivery, or implementation factors in addition to medications are becoming increasingly common in health sciences. A natural experimental approach to developing and refining such multicomponent interventions is to start with a large number of potential components and screen out the least active ones (e.g. the screening phase of MOST). Factorial designs can be used efficiently in this endeavor. In this chapter, we address the common criticisms and misconceptions regarding the use of factorial designs in the context of screening studies. We also provide an operationalization of screening studies. As an example we consider the use of a screening study in the development of a multicomponent smoking cessation intervention. Simulation results are provided to support the discussions.

3.1 Introduction

As discussed in Chapter I, multicomponent interventions composed of behavioral, delivery, or implementation factors in addition to medications are becoming increasingly common in health sciences. As has been recognized in the literature

[1, 18, 39, 86, 103], development and evaluation of these multicomponent interventions pose additional design challenges over those of single-component interventions, and these challenges tend to be addressed poorly by the standard two-group randomized controlled trials (often followed by non-experimental *post hoc* analysis to answer questions regarding individual components and their interactions). In particular one important challenge is to whittle down a large list of potential components, by screening out the least active components. Factorial designs are ideally suited to this endeavor [9].

The primary goal of this chapter is to consider the use of full and fractional factorial designs in screening out inactive components so as to aid in the development of high quality multicomponent interventions. We discuss how many of the criticisms prevalent in the literature concerning the use of full and fractional factorial designs no longer hold or are of lesser importance in *screening*[1] trials. A secondary goal is to provide an operationalization of screening trials using full and fractional factorial designs.

The present work is motivated by our participation in the design of a web-based smoking cessation study called *Project Quit* [88] that utilized fractional factorials. For illustrative purposes, we present a slightly modified version of *Project Quit*, following [62]. The investigators decided to study six components *Depth of outcome expectations*, *Depth of efficacy expectations*, *Depth of success stories*, *Personalization of message source*, *Mode of message framing*, and *Exposure schedule* (depth refers to the degree to which the communication was tailored to the background information on each individual). Since varying all six components across all possible levels in a single study was logistically prohibitive, the investigators decided to move forward

[1] Here the term *screening* refers to screening of intervention components, not screening of study participants.

in phases, where results of the research conducted in the first phase would inform the second, and so on [25, 62]. The goal of the first phase was to identify the active components and screen out inactive components. Each component was varied at two levels, as is common in screening studies. In addition all individuals were provided a 10-week free supply of nicotine patches. The investigators decided to use a 16-cell fractional factorial design (see section 3.3 for details). The primary outcome was self-reported seven-day point-prevalence abstinence at the 6-month follow-up from the date of randomization. More information on this study can be found in [88] and [62], and also in Chapter II of this dissertation.

The remainder of this chapter is organized as follows. Section 3.2 addresses common criticisms against the use of full and fractional factorial designs for developing multicomponent interventions. We provide an operationalization of the screening trials in section 3.3. Examples of possible follow-up studies are given in section 3.4. We conclude with an overall discussion in section 3.5. Technical review material on fractional factorial designs appears in the appendix (at the end of this chapter).

3.2 Factorial Designs for Screening Studies

Factorial designs were originally developed in the context of agricultural experiments [108, 36] and are now used in other areas including engineering [9, 107] and marketing research [29]. Their use in the medical and behavioral fields has been limited; however there have been a number of papers discussing the usefulness of these designs in medication and intervention trials [16, 13, 14, 103, 102, 23].

Prior to discussing common criticisms and concerns, we provide a brief review of both the design and analysis of screening studies. In screening, two-level factorial

designs, where all components are studied at two levels (these levels can be either present vs. absent, or two ethically acceptable doses of the component), are usually used since the goal is to identify important components rather than identify the optimal level of each component. If a two-level factorial design involves k components, then the total number of treatment combinations studied is 2^k. Each of the 2^k cells in the design corresponds to a group of subjects assigned to a particular treatment combination. In screening experiments, k is often large, rendering a full factorial design with 2^k cells infeasible. In such cases, fractional factorial designs (FFDs) [11] offer a useful alternative since they use fewer cells (see below for more discussion). For example, in the *Project Quit* study, a full factorial design with six components would need $2^6 = 64$ cells. But by using an FFD, it was possible to restrict the study to only 16 cells, and still be able to estimate all the main effects and some two-way interactions under reasonable assumptions.

In case of a continuous outcome, the analysis of a 2^k full factorial design (or a 2^{k-p} FFD) with total sample size n can be done using a linear regression model. One can use a model of the form $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{Y}$ represents the $n \times 1$ vector of observations on the outcome, $\mathbf{X}$ is the $n \times m$ design matrix, β is a $m \times 1$ vector of unknown parameters ($m = 2^k$ for full factorial and $m = 2^{k-p}$ for FFD, with more parameters if baseline variables are included in the analysis model), and ε is the $n \times 1$ vector of errors. It is assumed that $E(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2\mathbf{I}$. The design matrix consists of an intercept column, plus columns corresponding to each component and their interactions of different order coded in $-1/1$ (i.e., different factorial effects). The least squares estimator for β along with its covariance matrix are given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \quad Cov(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Note that the estimates of usual ANOVA quantities of interest like the main effect of a

component or the interaction between two or more components are directly related to the least squares regression estimates $\hat{\beta}$, provided the design matrix is coded in $-1/1$. As discussed by Byar et al. [15], the main effect of a component A_1 is estimated by $2\hat{\beta}_{A_1}$, A_1A_2 interaction is estimated by $4\hat{\beta}_{A_1A_2}$, and so on. In general, a p-component $(1 \leq p \leq k)$ interaction, say $A_{i_1} \ldots A_{i_p}$ (with $1 \leq i_1, \ldots, i_p \leq k$), is estimated by $2^p\hat{\beta}_{A_{i_1}\ldots A_{i_p}}$. If variance heterogeneity across different cells is anticipated in a study, one can use a robust estimator, e.g. *sandwich estimator* [104] of the covariance matrix given by

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(diag(\mathbf{Y} - \mathbf{X}\hat{\beta}))^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

in the linear regression model. But sample sizes should not be too small for this estimator to work well. Wu and Hamda [107] provide alternative methods to deal with variance heterogeneity. As discussed by Montgomery et al. [53], the regression approach can be used for *unbalanced*[2] data, and can estimate the factorial effects controlling for baseline or stratification variables. In case of binary (more generally, categorical) outcomes, the regression approach can be generalized via a generalized linear model. For example, if the outcome is binary, a logistic regression model can be used to analyze the data from a factorial design [107, Ch. 13]. See [62] and [88] for examples of such analyses.

3.2.1 Addressing Criticisms against Factorial Designs

Within the biostatistics literature, factorial designs have been assessed primarily in the context of medication trials; the objective has been to *evaluate* the usefulness of a combined medication over a single medication. In contrast, our objective is to

[2] *Balance* means that each level of each component appears in same number of cells and is assigned to the same number of subjects.

screen out inactive components of a multicomponent intervention; thus full and fractional factorial designs play a different role from that of evaluation. In this context, many of the common concerns re cost, feasibility, ethics, toxicity of combination drugs, interpretation of main effects in presence of active interactions, and concern about power for detecting interactions become moot or of lesser importance. Indeed many different complaints against factorial designs stem from a few fundamental issues and hence can be categorized as follows:

1. There are attractive alternatives to FFD.
2. It is not feasible to simultaneously implement multiple multicomponent interventions.
3. Some components cannot be crossed due to toxicity or ethical considerations.
4. The interpretation of main effects when interactions exist is complicated.
5. Power is low, or in other words the required sample size is high.

In the following, we address these broad classes of criticisms against factorial designs in the present context of screening studies for developing multicomponent interventions.

Attractive alternatives to FFD

The traditional approach of empirically developing multicomponent interventions, sometimes called the *treatment package strategy* [47, 103, 102], is to formulate a "likely best" intervention based on existing literature, theory, and clinical experience. Additionally investigators may use information from limited experimentation with some of the components either in stand-alone trials or in trials in which one

component is varied at a time while the remaining components are set at fixed levels. Implicitly one often assumes that more treatment is always better so the "likely best" intervention includes many components. An additional implicit assumption is that any ill effects due to including inactive components are minor. The developed multicomponent intervention is then evaluated in a standard two-arm randomized trial. These two-arm trials are confirmatory in that they are designed to provide high quality information on whether the multicomponent intervention performs better than the standard; they are not designed to provide direct information on which components are active, whether they have been set at optimal levels, and whether there is any interaction between the components [86]. To address the latter questions, investigators may use observational analyses, such as a dose-response with the level of subject adherence to the treatment as the dose [68, 1, 96, 97], or theory-based mediational analysis [103, 106]. The intervention is often refined based on the findings of these analyses, and then the refined version is tested in another two-arm randomized trial. Sometimes several such iterations are performed to refine the multicomponent intervention.

The main problem with this approach is that it depends heavily on the non-experimental, observational analyses. As is well-known [44, 72, 74], findings that are not based on randomization are contaminated by the likely presence of unknown *confounders*[3] (e.g. the variables that affect both the receipt of a component and the outcome). As a consequence, the effects of individual components and interactions may be misinterpreted resulting in a suboptimal intervention. Collins et al. [22] provided a head-to-head comparison between the above approach and an experimental procedure using FFDs in an extensive simulation study. This comparison was based

[3]In the literature on FFDs, the term *confounding* often refers to *aliasing* of effects. Here we use *confounding* to mean mixing of treatment effects with effects of other variables that affect both the receipt of treatment and the outcome, and thus keep *confounding* distinct from *aliasing*.

on a simulated model involving five components, varying levels of adherence to each component, an unknown confounder, and a continuous outcome. Also, the model included an antagonistic interaction between two components. The simulation results showed that the FFD-based experimental approach outperformed the traditional approach (two-arm randomized trial followed by observational analyses) in terms of various criteria, e.g. optimizing the mean outcome of the final intervention and identifying the best multicomponent intervention. See chapter IV for further details. Of course the relative merit of the FFD-based experimental approach depends on the degree of confounding; using observational analyses to investigate interactions might work well when the unknown confounder is only weakly related to the receipt of the components or the outcome.

Another alternative to FFDs is to conduct a series of *dismantling* or *subtractive* trials [103] where a "more complete" version of the multicomponent intervention is compared with a reduced version with one or more components eliminated. A close variant of this is known as the *constructive strategy* [103] or *treatment augmentation design* [45], where a base intervention is compared with an augmented version in which one or more components are added to the base intervention. Yet another alternative is known as the *comparative treatment strategy* [103], where several versions of the intervention are directly compared. For example, if there are k components under consideration, a comparative strategy would compare $(k + 1)$ experimental arms: k arms, each setting a single component at the high level and the rest at the low level, plus a control arm where all components are set at the low level. The above three approaches (i.e., dismantling, constructive, and comparative strategies) sometimes come under the umbrella term of single-factor designs [23] whenever the experimental arms under comparison differ by manipulating a single factor.

Note that there are several problems with using a series of single-factor experimental designs to construct a multicomponent intervention. First, as discussed by Box et al. [9, pp. 510-513], the use of single-factor designs often tacitly assumes that the effect of one component is independent of the levels of other components. This is not true in general, e.g. when there is a sizeable interaction between the components. Thus adopting a single-factor design often implicitly assumes that there is no interaction. Because of this limitation, using a series of single-factor designs to construct a multicomponent intervention may fail to achieve the best intervention.

The second problem regarding single-factor designs arises in designing the trials, e.g. deciding which factor to add (in constructive strategy) or subtract (in dismantling strategy), or which two versions of the multicomponent interventions to compare (in comparative strategy). These decisions are often driven by theory, cost, burden, or the results of observational analyses. To the extent that the results are driven by observational dose-response analyses of the amount of treatment received, they are vulnerable to confounding bias. As a consequence, in the sequence of single-factor trials conducted to find the best multicomponent intervention, active components may be accidentally eliminated in a dismantling strategy, and less active components may be erroneously added earlier than more active components in a constructive strategy.

A third problem with single-factor designs is that they often require many more subjects than comparable factorial designs to achieve similar power [23], rendering factorial designs a more efficient choice.

To summarize, in contrast to the treatment package strategy or the single-factor designs, inference about individual components in FFDs are strictly based on randomization, and hence less vulnerable to confounding bias. Furthermore, single-

factor experiments are not equipped to take care of interactions, and often have higher sample size requirement. Although some *aliasing* of effects happens in FFDs, the investigator can control this based on prior substantive knowledge (see below for more discussion on aliasing). Thus by using FFDs, one often trades uncontrolled confounding for controlled aliasing. Thus full and fractional factorial experimental designs offer a gold standard for developing multicomponent interventions. In the following, we will discuss feasibility.

Feasibility of the design

When the number of components (k) is moderately large, full factorial designs may be impractical due to cost of designing and implementing too many cells, i.e., making each treatment combination work together and ensuring implementation fidelity by staff [103]. This criticism has been the main motivation behind the development of FFDs. It is possible to select an FFD with substantially fewer cells, but still estimate the main effects (and sometimes important two-way interactions) without bias and with the same precision as in a full factorial design under plausible assumptions. A full factorial design allows the estimation of every individual factorial effect, including all higher order interactions. However, in absence of compelling prior theory or evidence to the contrary, third- and higher-order interactions are likely negligible in size in most multicomponent interventions [9, 62]. FFDs sacrifice the ability to estimate some of these higher order interactions, and in return, enable the study to have fewer cells. The choice of interactions to be sacrificed is informed by scientific theory, past studies, and investigator's experience. The practical price paid to buy the economy offered by an FFD is that the effects of interest, such as the main effects and two-way interactions, are *aliased* with some higher order interactions.

When two or more effects are aliased, one can estimate only the sum of the aliased effects. To overcome this problem, ideally an FFD is chosen in which each "aliased bundle" includes only one effect that is a priori believed to be active, with any other effects included in the bundle likely negligible in size. If this is not possible, follow-up experiments [52, 107, 25] can be conducted to settle any ambiguity about which effects are most important in the aliased bundle of effects. The above ideas were used by both *Project Quit* [88] and *Guide to Decide* [62] to design FFD trials. A technical review of aliasing and FFDs is given in Appendix A.

The strong use of theory and investigators' experience in determining which interactions to alias in an FFD is often initially disconcerting to scientists. Note however that in a two-arm randomized trial of a multicomponent intervention vs. control, the multicomponent intervention must be determined completely by theory and investigator's experience, and furthermore in these two-arm trials every factorial effect (main effects and interactions) is aliased with every other effect. Thus all analyses concerning individual components hinge on the use of a correct model; if the model is too simple then finding out what each effect is estimating is often difficult or even impossible. In this regard, FFDs offer a clearly better option in that the entire aliasing pattern is under the investigator's control, and there are principled ways (e.g. follow-up experiments) to disentangle any aliased effect. Moreover in non-experimental studies (that often follows the two-group comparisons) in which often the receipt of treatment depends on adherence to or availability of certain components, staff decisions as to who to offer what treatment etc., the resulting confounding is uncontrolled.

Often concerns about feasibility are intertwined with a perceived need to include many subjects in each cell of the design; this may occur because investigators er-

roneously think that comparisons between individual cells will be required. This however is not the case; see below for a discussion of this along with power considerations. None-the-less there are some situations in which investigators are unable to hire sufficient staff so as to implement multiple multicomponent interventions or are unable to train the staff to implement multiple multicomponent interventions simultaneously. In these settings FFDs are not feasible.

Inability to cross some components

To use factorial designs, one must be able to cross the components without changing dose (i.e., all combinations should be implementable). This has been a fundamental concern regarding the use of factorial designs in medication trials. In medication trials, toxicity often precludes the combined use of multiple components (e.g. drugs) unless the dosage is altered [16, 67]. That is, the combination of drug A and drug B uses lower doses of both A and B, compared to the case when either drug A or drug B is used alone. In such cases, the components lose their meanings, and factorial designs become inappropriate. Here we consider only those components, that when crossed, retain their meaning. This includes most behavioral, delivery, or implementation components, as well as multiple medications as long as they use different biological pathways.

When some components cannot be crossed, the clinical trials literature provides some approaches. Byar et al. [15] discussed *incomplete factorial* designs along with analysis strategies to take care of such cases. These designs are full or fractional factorial designs, minus some unpermitted combinations. Although these designs are not balanced (see the second footnote for a definition of balance), one can still estimate many of the relevant factorial effects.

Interpretation of main effects

It is well-known that the definition of the main effect, in the presence of sizeable interactions [16, 67], differs from investigators' conceptual definition of the effect of a component. To address this criticism, here we provide precise definitions of main effects and simple effects that commonly arise in various designs for multicomponent interventions, and establish their interrelationship.

For simplicity, consider a 2×2 factorial design with two components, say A_1 and A_2, and continuous outcome Y. The presence and absence (or, high and low level) of each component is coded $+1$ and -1 respectively. Let $\mu_{(-,-)}$, $\mu_{(-,+)}$, $\mu_{(+,-)}$ and $\mu_{(+,+)}$ be the mean outcomes corresponding to the absent-absent, absent-present, present-absent, and present-present cells of the design respectively. At the population level, the main effect of the component A_1 is defined by $\frac{1}{2}(E_1 + E_2)$, where $E_1 = (\mu_{(+,+)} - \mu_{(-,+)})$ and $E_2 = (\mu_{(+,-)} - \mu_{(-,-)})$ are two *simple effects*, denoting the effect of A_1 when A_2 is fixed at high and low level respectively. Thus the main effect of A_1 is defined as the average[4] of the two simple effects E_1 and E_2, and hence can be interpreted as the effect of A_1 when half the subjects in the population are exposed to (the high level of) A_2 and the remaining half are not. On the other hand, when conceptualizing the treatment effect of A_1, an investigator usually thinks of the simple effect denoting "the effect of A_1 in absence of A_2" [67, p. 506], i.e., E_2. In absence of interaction between the components, this mismatch does not cause a problem since the two simple effects are equal. However, the main effect could be very different from the simple effect of A_1 in presence of a sizeable interaction.

[4]The main effect is usually defined as the average of simple effects as presented above. However, in general, a weighted average over the distribution of the simple effects in the population may be a better definition.

If a dismantling strategy is followed (dismantling A_1 from the full package involving both A_1 and A_2), then the effect estimated is simply $(\mu_{(+,+)} - \mu_{(-,+)}) = E_1$. This effect could also be estimated if the constructive strategy is followed (augmenting A_1 to the base intervention consisting of A_2 only). Thus these alternative designs estimate simple effects rather than main effects. Lastly one can imagine a *treatment package effect*, e.g. $(\mu_{(+,+)} - \mu_{(-,-)})$, which is estimated when the "likely best" package consisting of the present or high level of all the components is compared with a control consisting of the absent or low levels of all the components. This does not correspond to any of the simple effects.

For three two-level components, the main effect of a component A_1 is defined as $\frac{1}{4}(E_1 + E_2 + E_3 + E_4)$, where $E_1 = (\mu_{+,+,+)} - \mu_{(-,+,+)})$, $E_2 = (\mu_{(+,+,-)} - \mu_{(-,+,-)})$, $E_3 = (\mu_{(+,-,+)} - \mu_{(-,-,+)})$, and $E_4 = (\mu_{(+,-,-)} - \mu_{(-,-,-)})$ are the four simple effects (and also can be interpreted as the effects resulting from different dismantling or constructive trials). The most common simple effect is E_4, i.e., "effect of A_1 in absence of other components", and is often conceptualized as the treatment effect of A_1. In general, for a setting involving k two-level components, there are 2^{k-1} simple effects that can be interpreted as effects resulting from different constructive or dismantling trials. The main effect is simply the average of these 2^{k-1} simple effects.

To more clearly understand the alternative definitions and how they differ in the presence of an interaction, consider a regression formulation. Suppose the true data-generating model, where A_1, A_2 are coded in 0/1, is given by

$$Y = b_0 + b_1 A_1 + b_2 A_2 + b_{12} A_1 A_2 + \varepsilon. \tag{3.1}$$

If we use a regression analysis with the $-1/1$ coding, e.g. we fit $\beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_{12} A_1 A_2$, then we estimate the following transformed model (now A_1, A_2 are coded

in $-1/1$):

$$Y = \Big(b_0 + \frac{b_1 + b_2}{2} + \frac{b_{12}}{4}\Big) + \Big(\frac{b_1}{2} + \frac{b_{12}}{4}\Big)A_1 + \Big(\frac{b_2}{2} + \frac{b_{12}}{4}\Big)A_2 + \Big(\frac{b_{12}}{4}\Big)A_1A_2 + \varepsilon.$$

The main effect of A_1 is $2\hat{\beta}_1$, which estimates the population quantity $2\Big(\frac{b_1}{2} + \frac{b_{12}}{4}\Big) = b_1 + \frac{b_{12}}{2}$ (this main effect continues to be the average effect of A_1 on Y over the levels of A_2). In contrast, the two simple effects of A_1 are b_1 (effect of A_1 when A_2 is absent), and $b_1 + b_{12}$ (effect of A_1 when A_2 is present). The main effect and the effect commonly conceptualized as the treatment effect of A_1, i.e., b_1, differ by the quantity $\frac{1}{2}b_{12}$ in presence of an active interaction ($b_{12} \neq 0$). If we apply the reasoning of the *Hierarchical Ordering Principle*[5] [107] to this setting, then in general we expect that b_{12}, if nonzero, is likely of smaller size than b_1 and b_2.

To summarize, when there is an interaction, the main effect has the interpretation of the average effect of A_1 on Y over the levels of other components. This is quite different from what is often conceptualized as the treatment effect of A_1, e.g. the simple effect of A_1 on Y setting other components to lower level. However the crucial point is that in screening studies, the goal is to screen components efficiently, and not to estimate either the simple effect or the main effect of a component per se. The important issue for screening is whether this difference in definition impinges on our ability to screen components. So in this context, the concern about the definition of main effects is actually a concern about power to screen components. We address this concern below in great detail (see the third issue below under the Power heading).

[5]The *Hierarchical Ordering Principle* is an assumption commonly made in design of experiments in the absence of substantive theory or prior results suggesting otherwise. It states that lower-order effects are more likely to be important than higher-order effects, and effects of the same order are equally likely to be important.

Power

Several issues lead to concerns about power when factorial designs are considered. First, investigators sometimes use factorial designs to evaluate or compare a few multicomponent interventions, e.g. compare one cell against another cell [42], or otherwise assess simple effects. This naturally leads to a large sample size requirement since each cell (group of subjects) must be large. However to screen components, we primarily focus on main effects and sometimes also a few anticipated two-way interactions. The focus on main effects and lower order interactions for the purpose of screening can be partially justified by the *Hierarchical Ordering Principle* [107], which says that main effects and lower order interactions are likely more important than higher order interactions. Recall the main effect of a factor is an average of all the 2^{k-1} simple effects. Thus even though several components are studied, the total sample size required for assessing the significance of a main effect is the same as that for a two-group trial involving a single component (for example in a linear model, the estimator of the main effect is proportional to the difference between the means of two groups of cells; all cells in the FFD belong to one or the other group). Furthermore, in the multiphase approach to intervention development [25, 62], ascertaining the best treatment combination is done through follow-up studies, in which one usually focuses on only a few combinations of components while holding the levels of the remaining components constant. See section 3.4 for a discussion of follow-up studies.

Second, there is concern about the loss of balance and subsequent loss of power due to study dropout. In most intervention studies, patient dropout is inevitable, thus resulting in unequal cell sizes. As discussed by Montgomery et al. [53], this is an issue for all clinical trials rather than a criticism of factorial designs; modern-day missing data techniques will be needed in the analysis, as is the case with any

randomized clinical trial.

A third issue related to power is how one should formulate the test statistics to detect the effects of treatment components in a screening study. Note that in a screening study the goal is to screen out inactive components, and not to estimate either a simple effect or a main effect per se. Below we show that even when the data are generated using non-zero simple effects, often the power to detect the resulting main effect is higher than the power to detect the original simple effect. Hence in a screening study, formulating the test statistics based on main effects is in general better than formulating test statistics based on simple effects. To discuss this consider again the 2×2 factorial design with two components, say A_1 and A_2, r subjects per cell, and the continuous outcome Y. The true data-generating model is specified in terms of simple effects, which is consistent with an investigator's conceptualization. Thus the true data-generating model is given by (3.1) in which the components A_1, A_2 are coded in 0/1. In the following, we show that by basing the test statistic on main effects, we can in general screen non-zero simple effects with greater power.

For simplicity, assume $\text{Var}(\varepsilon) = \sigma^2$ is known (and homogeneous across cells). If a linear regression model with 0/1 coding is used as in Piantadosi [67, pp. 508-509] then the following model is fit:

$$\beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_{12} A_1 A_2. \tag{3.2}$$

Here β_1, the coefficient of A_1, is a simple effect representing the comparison of the $(1,0)$ cell with the $(0,0)$ cell, i.e., $\beta_1 = \mu_{(1,0)} - \mu_{(0,0)} = b_1$, where $\mu_{(1,0)}$ is the population mean of Y in the $(1,0)$ cell, and so on. Now β_1 is estimated by $\hat{\beta}_1 = \bar{Y}_{(1,0)} - \bar{Y}_{(0,0)}$, where $\bar{Y}_{(1,0)}$ is the sample mean of Y in the $(1,0)$ cell, and so on. Clearly, $E(\hat{\beta}_1) = b_1$,

and

$$Var(\hat{\beta}_1) = Var(\bar{Y}_{(1,0)}) + Var(\bar{Y}_{(0,0)}) = \frac{\sigma^2}{r} + \frac{\sigma^2}{r} = \frac{2\sigma^2}{r}.$$

So the signal-to-noise ratio (SNR) governing the power to screen A_1 when basing the test statistics on simple effects is

$$SNR_{\text{simple}} = \frac{|\mathrm{E}(\hat{\beta}_1)|}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} = \frac{|b_1|\sqrt{r}}{\sqrt{2}\sigma}.$$

On the other hand, if we use the analysis model (3.2) with $-1/1$ coding, it follows that

$$\begin{aligned} \beta_1 &= \frac{1}{4}\Big[(\mu_{(+,+)} - \mu_{(-,+)}) + (\mu_{(+,-)} - \mu_{(-,-)})\Big] \\ &= \frac{1}{2} \times (\text{the main effect of } A_1) \\ &= \frac{1}{2} \times (\text{the average of 2 simple effects}), \end{aligned}$$

and is estimated by the sample version $\hat{\beta}_1$ (where μ is replaced by $\bar{Y}$). Then,

$$\begin{aligned} E(\hat{\beta}_1) &= \beta_1 = \Big(\frac{b_1}{2} + \frac{b_{12}}{4}\Big), \\ Var(\hat{\beta}_1) &= \frac{1}{4} \times \frac{1}{2} \times (\text{variance of an estimated simple effect}) = \frac{1}{4} \times \frac{1}{2} \times \frac{2\sigma^2}{r} = \frac{\sigma^2}{4r}. \end{aligned}$$

So the signal-to-noise ratio governing the power to screen A_1 when basing the test statistics on main effects is

$$SNR_{\text{main}} = \frac{|\mathrm{E}(\hat{\beta}_1)|}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} = \Big|b_1 + \frac{b_{12}}{2}\Big|\frac{\sqrt{r}}{\sigma}.$$

A measure of relative efficiency of the two ways of forming the test statistics in screening A_1 is given by

$$\eta = \frac{SNR_{\text{main}}}{SNR_{\text{simple}}} = \sqrt{2}\Big|b_1 + \frac{b_{12}}{2}\Big|/|b_1| = \sqrt{2}\Big|1 + \frac{b_{12}}{2b_1}\Big|.$$

In absence of an interaction (i.e., $b_{12} = 0$), $\eta = \sqrt{2} > 1$, and hence basing the test statistic on main effects gives higher power for screening components. In case of synergistic interaction (i.e., b_1 and b_{12} are of same sign), η is even larger, leading to higher power. Even in case of antagonistic interaction (i.e., b_1 and b_{12} are of opposite sign), the way of basing the test statistic on main effects gives higher power in screening components (i.e., $\eta > 1$) if $b_1 < 0$ and $0 < b_{12} < -(2 + \sqrt{2})b_1$, or if $b_1 > 0$ and $0 > b_{12} > -(2 - \sqrt{2})b_1$. If we have k (≥ 2) components in a factorial experiment, and there may be a two-way but no higher-order interaction in the true data-generating model, then the relative efficiency of the two ways of forming the test statistic (measured by η) increases with k. A verification of this appears in Appendix B.

To illustrate the power implications of basing the test statistics on main effects rather than simple effects in a regression analysis, we consider a small simulation study with the data-generating model $Y|A_1, A_2 \sim N(\mu = b_0 + b_1A_1 + b_2A_2 + b_{12}A_1A_2, \sigma = 1)$, where A_1, A_2 are coded in 0/1. That is, the data-generating model is specified in terms of simple effects (as is usually conceptualized by an investigator). The coefficients b_1, b_2 are set according to Cohen's [21] small or medium effect size (i.e., $b_1 = b_2 = 0.2, 0.5$). The coefficient b_{12} of the interaction term is varied: $b_{12} = b_1, b_1/2, 0, -b_1/2, -b_1$ (i.e., same size and sign as b_1, half the size of and same sign as b_1, absent, half the size of b_1 but of opposite sign, same size as b_1 but of opposite sign). A 0.05 level of significance is used throughout, while varying the sample size: $n = 100, 200, 500$. The goal of this simulation is to illustrate that even when the data-generating model is specified in terms of simple effects, basing the test statistics on main effects leads to higher power in most settings than basing the test statistics on simple effects. Note that the signal-to-noise ratios govern the

corresponding powers. In the following, we consider the power to screen A_1 both in presence and absence of an interaction term A_1A_2 (synergistic as well as antagonistic). Table 3.1 contains a Monte Carlo estimate (using 1000 iterations) of the power for screening A_1 under different scenarios.

Table 3.1: Power to screen A_1 in absence and presence of an interaction

n	Interaction Size (b_{12})	Interaction Type	$b_1 = 0.2$		$b_1 = 0.5$	
			Using simple effect	Using main effect	Using simple effect	Using main effect
100	Same ($= b_1$)	synergistic	0.1030	0.2910	0.4150	0.9550
	Half ($= \frac{b_1}{2}$)	synergistic	0.1030	0.2290	0.4150	0.8730
	Absent ($= 0$)	none	0.1030	0.1720	0.4150	0.6830
	Half ($= -\frac{b_1}{2}$)	antagonistic	0.1030	0.1110	0.4150	0.4420
	Same ($= -b_1$)	antagonistic	0.1030	0.0820	0.4150	0.2290
200	Same ($= b_1$)	synergistic	0.1690	0.5440	0.6920	1.0000
	Half ($= \frac{b_1}{2}$)	synergistic	0.1690	0.3940	0.6920	0.9870
	Absent ($= 0$)	none	0.1690	0.2840	0.6920	0.9430
	Half ($= -\frac{b_1}{2}$)	antagonistic	0.1690	0.1720	0.6920	0.7510
	Same ($= -b_1$)	antagonistic	0.1690	0.1040	0.6920	0.3940
500	Same ($= b_1$)	synergistic	0.3460	0.9210	0.9740	1.0000
	Half ($= \frac{b_1}{2}$)	synergistic	0.3460	0.8040	0.9740	1.0000
	Absent ($= 0$)	none	0.3460	0.6050	0.9740	1.0000
	Half ($= -\frac{b_1}{2}$)	antagonistic	0.3460	0.3730	0.9740	0.9870
	Same ($= -b_1$)	antagonistic	0.3460	0.1890	0.9740	0.8040

Note that in Table 3.1, the power to screen A_1 is higher in general when the test statistic is based on main effects compared to when it is based on simple effects (e.g. comparing the 4th vs. 5th column, and comparing the 6th vs. 7th column), except when the interaction is of same size and opposite in sign as the simple effect of A_1 (as expected from the above discussion). However according to the *Hierarchical Ordering Principle* [107], interactions are usually of smaller order of magnitude than the main effects (absent strong scientific theory to the contrary), and hence this is a fairly unlikely scenario. A secondary point to note is that when the test statistic is based on main effects, there is a decrease in power to screen A_1 as the interaction term b_{12} decreases from highly synergistic to highly antagonistic (moving down the 5th and

7th columns). However, when the test statistic is based on simple effects, the power for screening A_1 is independent of the size of the interaction term b_{12} (moving down the 4th and 6th columns). But the decrease in power in the 5th and 7th columns due to interaction often does not pose a serious threat (as compared to the loss of power from using simple effects in the test statistic) if the goal is to screen components, since in most settings the way of basing the test statistic on main effects gives better power anyway.

A fourth issue related to power is the power to detect interactions. Factorial designs are often criticized on the ground that the power to detect an interaction is much lower than the power to detect a main effect of the same size [67, 53]. However, it is also recognized that factorial designs are the only experimental designs that can systematically investigate interactions. To overcome the low power for detecting interactions in a confirmatory (not screening) trial, the general recommendation in the literature [16] is that if an interaction is strongly anticipated based on the investigator's prior knowledge, the study should be powered with larger sample size. When criticizing factorial designs on the ground of low power for interactions in the present context of screening trials for developing multicomponent interventions, it is useful to consider the pros and cons of the possible alternatives. The natural alternative is to conduct non-experimental analyses using treatment adherence or other post-randomization outcomes as doses or factor levels from a randomized trial or to use observational data sets. As discussed previously, the relative merit of FFDs over the above strategy depends on the degree of confounding in the data. The crux is that the low power to detect interactions in a factorial design can be offset by its ability to perform valid estimation and inference, and its ability to control (by design) aliasing in a principled manner, in comparison to observational analyses.

3.3 Operationalization of Screening Trials

This section provides an example of how screening trials can be operationalized using FFDs. The choice of an appropriate FFD is often governed by prior knowledge regarding the intervention to be developed. To move forward, two definitions are useful. An FFD is completely characterized by its *defining relation* [107], a rule from which the aliasing pattern of the FFD can be obtained. Moreover, FFDs are sometimes categorized by their *resolution*. Loosely speaking, the higher the resolution, the better is the design. Resolution IV and resolution V designs are considered here. In particular, in a resolution V design, main effects are aliased with 4-way (or higher order) interactions, and 2-way interactions are aliased with 3-way (or higher order) interactions. Likewise in a resolution IV design, main effects are aliased with 3-way (or higher order) interactions, and 2-way interactions are aliased with other 2-ways (or higher order). Typically resolution V designs are better than resolution IV designs, but resolution V designs require more cells. Hence investigators may adopt lower resolution designs due to cost and feasibility constraints. Further review of the *defining relation* and *resolution* are given in the Appendix A. In the following, we first discuss the screening design used in the *Project Quit* study. Next, we discuss a general approach to construct screening designs (e.g., appropriate FFDs).

Screening design in the Project Quit study

Denote the six components of the *Project Quit* study, e.g. depth of outcome expectations, depth of efficacy expectations, depth of success stories, personalization of message source, mode of message framing, and exposure schedule by A_1, A_2,

A_3, A_4, A_5, and A_6 respectively. In this study, prior knowledge suggested that the interactions between outcome expectations and efficacy expectations (A_1A_2), outcome expectations and success stories (A_1A_3), outcome expectations and message framing (A_1A_5), and efficacy expectations and message framing (A_2A_5) were likely active (let us call them *anticipated* interactions), and that all other interactions should be negligibly small in size. So a design was constructed such that one could estimate the A_1A_2, A_1A_3, A_1A_5, and A_2A_5 interactions, assuming all others to be small. Due to cost constraints, 16 cells were used in the design. So the design used was a 16-cell FFD with the defining relation

$$I = A_1A_2A_4A_5 = A_1A_3A_4A_6 = A_2A_3A_5A_6. \tag{3.3}$$

This is a resolution IV design where some of the 2-way interactions are aliased with other 2-way interactions. The anticipated 2-way interactions are listed on the left-hand side of the following aliasing equations (obtained from (3.3)):

$$A_1A_2 = A_4A_5$$

$$A_1A_3 = A_4A_6$$

$$A_1A_5 = A_2A_4$$

$$A_2A_5 = A_1A_4$$

Note that the anticipated interactions were aliased with other 2-way interactions that were considered negligible, and hence were estimable without bias. The defining relation $I = A_1A_2A_3A_5 = A_1A_3A_4A_6 = A_2A_4A_5A_6$ was "cleverly" chosen to accomplish this goal. Of course, the investigator's assumption about the interactions could be wrong. But one can verify any critical working assumptions made in the screening study using follow-up studies [25].

Screening design construction in general

As a starting point we assume that regardless of the number of components studied, the number of cells used can be at most 16 (equal to the number of cells used in the *Project Quit* study). Of course this number can vary from one setting to another. If 4 or fewer components are to be studied, a full factorial design can be used. If 5 components, say $A_1, \ldots, A_5$ are to be studied, then one should use the resolution V FFD with the defining relation $I = A_1A_2A_3A_4A_5$ (this is the case in the *Guide to Decide* project described by [62]). If 6 components, say $A_1, \ldots, A_6$ are to be studied, resolution IV designs are generally recommended. If prior knowledge suggests a few anticipated 2-way interactions, an FFD can be chosen carefully so that the anticipated 2-way interactions are not aliased with each other (this consideration often drives the construction of the design). Assuming the unanticipated interactions to be negligible, this ensures that each anticipated interaction can be estimated without bias. When there is only one anticipated interaction, any 16-cell resolution IV FFD can be used. However, for two or more anticipated interactions, choices are limited. Software (e.g., SAS PROC FACTEX, JMP, Minitab) can be used to generate the designs in such cases (they provide one possible design that satisfies the constraints of resolution and/or anticipated interactions, instead of giving the complete list of possible designs). For two or three anticipated interactions, the complete set of recommended designs are given in Table 3.2.

Power and sample size in screening trials

In a screening trial using a factorial design, the power calculation used to size the trial focuses on main effects of each component. Thus the power calculation is similar to that of a two-arm randomized trial in that the two levels of a single

component (averaged over the levels of all other components) serve as the two arms. Below we provide the power calculation for the Project Quit study as an example. For Project Quit, the planned initial recruitment size was 2000; this number was chosen to achieve a total sample size of 1500 for the analysis, anticipating a 75% response rate at the 6-month follow-up. Assuming no differential attrition across cells, this meant roughly 750 subjects per level of each intervention component. The primary outcome was binary, e.g. seven-day point-prevalence smoking cessation at the 6-month follow-up. So the power analysis involved binomial calculations (using a normal approximation) assuming a baseline average cessation rate of 10% found in a previous study [33]. For each main effect, the sample size of 750 per level provides approximately 80% power for detecting a 4.5% difference in cessation rates. The same power characteristics exist for each of the six components. Note that to

Table 3.2: Recommended resolution IV FFDs under varying anticipated interactions

Case	Anticipated interactions of the form	Recommended designs (defining relations)
1	A_1A_2, A_3A_4 (no component shared)	$I = A_1A_2A_3A_5 = A_1A_3A_4A_6 = A_2A_4A_5A_6$
		$I = A_1A_2A_3A_5 = A_2A_3A_4A_6 = A_1A_4A_5A_6$
		$I = A_1A_2A_4A_5 = A_1A_3A_4A_6 = A_2A_3A_5A_6$
		$I = A_1A_2A_4A_5 = A_2A_3A_4A_6 = A_1A_3A_5A_6$
		$I = A_1A_2A_3A_6 = A_1A_3A_4A_5 = A_2A_4A_5A_6$
		$I = A_1A_2A_3A_6 = A_2A_3A_4A_5 = A_1A_4A_5A_6$
		$I = A_1A_2A_4A_6 = A_1A_3A_4A_5 = A_2A_3A_5A_6$
		$I = A_1A_2A_4A_6 = A_2A_3A_4A_5 = A_1A_3A_5A_6$
2	A_1A_2, A_1A_3 (one component shared)	$I = A_1A_2A_4A_5 = A_1A_3A_4A_6 = A_2A_3A_5A_6$
		$I = A_1A_2A_4A_6 = A_1A_3A_4A_5 = A_2A_3A_5A_6$
		$I = A_1A_2A_4A_5 = A_1A_3A_5A_6 = A_2A_3A_4A_6$
		$I = A_1A_2A_5A_6 = A_1A_3A_4A_5 = A_2A_3A_4A_6$
		$I = A_1A_2A_4A_6 = A_1A_3A_5A_6 = A_2A_3A_4A_5$
		$I = A_1A_2A_5A_6 = A_1A_3A_4A_6 = A_2A_3A_4A_5$
3	A_1A_2, A_3A_4, A_5A_6	same as case 1
4	A_1A_2, A_1A_3, A_4A_5	$I = A_1A_2A_5A_6 = A_1A_3A_4A_6 = A_2A_3A_4A_5$
		$I = A_1A_2A_4A_6 = A_1A_3A_5A_6 = A_2A_3A_4A_5$
5	A_1A_2, A_1A_3, A_2A_4	$I = A_1A_2A_5A_6 = A_1A_3A_4A_5 = A_2A_3A_4A_6$
		$I = A_1A_2A_5A_6 = A_1A_3A_4A_6 = A_2A_3A_4A_5$
6	A_1A_2, A_1A_3, A_1A_4	$I = A_1A_2A_5A_6 = A_1A_3A_4A_5 = A_2A_3A_4A_6$
7	A_1A_2, A_1A_3, A_2A_3	same as case 2

achieve the same power to detect the same difference in cessation rates, one would need the same sample size in a usual two-arm study (so the sample size requirement is not increased by using a factorial design). The formula for calculating power in the present set-up is given by

$$\Phi\left(\frac{\sqrt{\frac{n}{2}}|\Delta| - z_{\alpha/2}\sqrt{2p(1-p)}}{\sqrt{p(1-p)+(p+\Delta)(1-p-\Delta)}}\right),$$

where n is the total sample size, p is the baseline cessation rate, Δ is the change in cessation rate to be detected, α is the Type I error, $z_{\alpha/2}$ is the upper $100(\frac{\alpha}{2})\%$ cutoff point of a standard normal distribution, and Φ is the standard normal distribution function.

Additional practical considerations regarding study duration and cost

A primary advantage of using factorial designs in a screening study lies in its efficiency, i.e., its ability to answer several screening questions (regarding multiple intervention components) quickly from a single study. The use of an FFD-based approach in Project Quit was motivated by the concern that advances in communication technologies were moving well beyond the understanding of message content, presentation, and delivery principles in the field of smoking cessation. Investigators of this study realized that research using the field's most widely-used designs (e.g. randomized trials with a small number of groups) [33, 87, 89, 91] would take years to assess even a few basic questions. By the time these findings would be disseminated, the technology and target populations would likely have been changed (e.g. become more sophisticated in their understanding of a communications channel), and consequently the field would continue to lag behind. Thus in the context of this concern, the FFD-based multiphase approach provided a huge benefit by offering a shorter total study duration to answer so many questions compared to the alternative

designs.

There are two kinds of cost associated with designing multicomponent intervention trials, e.g., (1) cost associated with sample size requirement, and (2) cost of designing and implementing different cells. We have already discussed that the sample size requirement does not go up by using an FFD. The only additional cost of designing an FFD over a two-group trial is the cost of designing and delivering too many versions of the intervention which might limit the applicability of FFDs in certain settings. In case of Project Quit, the intervention was delivered entirely through the internet. So the delivery of 16 versions of the multicomponent intervention did not cost additional staff time and training over and above the cost of software programming to generate the different versions, which turned out to be manageable. See Collins et al. [23] for a detailed comparison of FFDs with single-factor designs (dismantling, constructive and comparative trials) from a resource management perspective.

Screening analysis

The screening analysis uses a linear model (in case of continuous outcome) or a generalized linear model (in case of binary or categorical outcome). A few considerations to be made during the analysis are:

1. The level of significance α for testing the effects in the screening study might be set higher than 0.05 to achieve greater power for detecting effects. α can be viewed as a tuning parameter of the procedure. One possible choice is to use $\alpha = 0.1$ for the main effects and *anticipated* two-way interactions, and a Bonferroni-corrected 0.1 level for the *unanticipated* interactions.

2. As an alternative (or augmentation) to performing significance tests at the

screening study, one can rank-order the absolute values of the test statistic corresponding to the factorial effects (or equivalently p-values) and move to follow-up studies with the largest m. Then this m becomes a tuning parameter of the procedure. This approach should work better in case all individual effects are small, but together they produce some effect (significance test often accepts the null hypothesis of no effect in such cases, and hence perform poorly). To be resistant to the noise in the data, one may choose to rank-order only the main effects and *anticipated* interactions. This strategy with $m = 3$ was followed in the simulation study described in [22].

Examples of the screening analysis in the *Guide to Decide* and *Project Quit* studies can be found in [62] and [88]. Based on the screening analysis of the *Project Quit* study, the investigators decided to move to the follow-up study with the components having the highest two p-values (e.g. *success stories* and *message source*). Furthermore since three of the components (outcome expectations, efficacy expectations and success stories) were set at levels corresponding to high depth of tailoring versus low depth of tailoring, the investigators considered a regression of overall depth of tailoring (over all components) and found that as the depth of tailoring increased the smoking cessation rate increased. Hence the investigators decided to use a high depth of tailoring in the follow-up study.

3.4 Follow-up Studies

In the process of developing a multicomponent intervention, an investigator often conducts follow-up studies involving the *significant*[6] factorial effects from the screen-

[6]Throughout this section, we use the term *significant* loosely to mean any effects that come out important according to the screening analysis strategy outlined above.

ing study to fine-tune the results, e.g. finding the best level (or dose) of a significant component which is either continuous or has more than two levels by a dose-response experiment (where the subjects are randomized to ethically acceptable doses of the component), or de-aliasing significant aliased interactions by a smaller factorial experiment. In this section, first we provide a few hypothetical examples (of varying level of complexity) of follow-up studies to provide some general intuition, and then briefly describe the follow-up phase of the *Project Quit* study.

Hypothetical examples

In the following examples, for simplicity, we assume that there are 6 components in the study, e.g. $A_1, \ldots, A_6$, out of which only A_1 is a 3-level component (say, high, medium, low levels – only high and low levels are studied at the screening trial) and the rest are binary (high and low). High values of the outcome are preferred. A 16-cell resolution IV FFD is used as the screening design (see section 3.3 for details). We assume throughout that three-way (or higher-order) interactions are negligible in size compared to the noise in the data; hence even though main effects are aliased with three-way interactions, we assign the estimated effect to the main effect.

Example 1: Suppose the significant effects along with their signs based on screening analysis are:

$$A_1(+), A_2(+), A_3(-), A_5(-), A_2A_3 = A_4A_5(-),$$

where the aliased interaction A_2A_3 is unanticipated but A_4A_5 is anticipated. So the investigator may dismiss A_2A_3 as a possible effect and assign the observed effect entirely to A_4A_5. Since the main effect of A_4 is insignificant, the main effect of A_5 is negative, and the A_4A_5 interaction is negative, it is reasonable to set A_4 at its high

level and A_5 at its low level to maximize the mean outcome. Also, from the signs of the estimated main effects of A_2 and A_3 (and ignoring the unanticipated A_2A_3 interaction), A_2 and A_3 should be set at their high and low levels respectively. Since A_6 is insignificant, it should be set at low level. Hence the follow-up study might be a 2-group trial varying A_1 at its medium and high levels (since its main effect is positive), setting A_2, A_3, A_4, A_5, and A_6 at high, low, high, low, and low level respectively. If A_4 is an expensive or particularly component then it may be worthwhile to affirm that A_4 is not significant yet its interaction with A_5 is. In that case, the follow-up study can be a 8-group trial where the two levels of A_1 (high/medium) are crossed with two levels of A_4 and A_5 each. In all the groups, A_2, A_3, and A_6 should be set at high, low, and low level respectively.

Example 2: Suppose the significant effects along with their signs based on screening analysis are:

$$A_1(+), A_3(+), A_1A_3 = A_2A_6(-),$$

where aliased interaction A_1A_3 is anticipated but A_2A_6 is unanticipated. As before, we dismiss A_2A_6 based on prior considerations and assign the observed effect to A_1A_3. The follow-up study could be a 6-group trial crossing three levels of A_1 with two levels of A_3. In all the groups, the levels of A_2, A_4, A_5, and A_6 should be set at the low level.

Example 3: Suppose the significant effects along with their signs based on screening analysis are:

$$A_1(+), A_2(+), A_3(+), A_5(-), A_2A_3 = A_4A_5(-),$$

where both the interactions A_2A_3 and A_4A_5 involved in the aliased bundle are unan-

ticipated. Since the main effect of A_4 is insignificant, the main effect of A_5 is negative, and the aliased A_4A_5 interaction is negative (even though we are not sure if the observed effect is really due to A_4A_5), one reasonable step would be to set A_4 at its high level (provided the high level of A_4 is not very expensive or burdensome) and A_5 at its low level (note that our decision about the optimum levels of A_4 and A_5 would be same when A_4A_5 effect is really negative as when A_4A_5 is null). Also, we would set A_6 to the low level. If there is concern about the potential A_2A_3 interaction then the follow-up study could be a 8-group trial, where medium and high levels of A_1 are crossed with the two levels of A_2 and A_3 each to form the 8 groups (setting A_4, A_5, and A_6 at high, low, and low levels respectively).

Follow-up study design of Project Quit

An alternative to the follow-up studies outlined above is provided by *Project Quit* study, in which all components were two-level (hence a dose-response experiment was unnecessary) and no (unanticipated) aliased interaction were found significant (hence no de-aliasing experiment was necessary). The investigators decided to study different aspects (not studied in the screening trial) of the two important components (e.g. *success stories* and *message source*). The decision was to vary message source at two levels (high/low) of additional personalization, and to vary success stories in terms of the archetype (language and picture) of the hypothetical character in the story at three levels (e.g. a rebel, care-giver, or self-made character). Two new two-level components, e.g. *order* (of appearance on the web site: *success stories* first vs. *health advice* first) and *email quit status request* (yes/no) were added to the follow-up study. Subjects randomized to the "yes" level of *email quit status request* were contacted by the study staff at regular intervals about their quit status. The

follow-up study consisted of 25 groups in total: 24 groups from the $2 \times 3 \times 2 \times 2$ factorial structure of the above 4 components, plus a control group. In all groups, the original components from the screening trial not studied in the follow-up study were set as follows: deeply-tailored efficacy expectation and outcome expectation messages, gain framing, and multiple exposures. All three levels of the success stories were also deeply-tailored. The control group received the best intervention according to the results of screening study (e.g., highly personalized source at the first session only, deeply-tailored story with fixed archetype as in the screening study, deeply-tailored efficacy expectation and outcome expectation messages, gain framing, and multiple exposures) – they did not receive any email about their quit status.

3.5 Discussion

Multicomponent interventions are becoming increasingly common in health sciences. In this chapter, we have addressed the criticisms and misconceptions regarding the use of full and fractional factorial designs (e.g. attractive alternatives to and feasibility of such designs, inability to cross components, interpretation of main effects, and concerns about power) in the context of screening studies to develop multicomponent interventions. Other issues regarding the use of factorial designs, as discussed by Couper et al. [26], are slower recruitment rate (since subjects need to meet the inclusion criteria for all the components) and potential lower compliance (due to a more complicated treatment protocol) than single-component trials. However, these are common to any studies of multicomponent interventions, and not problems specific to factorial designs.

We provided some examples of follow-up studies that often need to be conducted

(e.g. to de-alias significant aliased interactions) after the completion of the screening study. Further strategies for conducting follow-up studies can be found, for example, in [52] and [107]. Also, in case there is at least one component with more than two levels (e.g. a continuous component), dose-response experiments [10, 61] where subjects are randomized to ethical doses should be used to find the optimal dose of these components. Operationalizing a wider variety of follow-up studies needs more targeted future research.

In our discussion of FFDs, we assumed that third- and higher-order interactions are negligible [9, 62]. This is not a binding constraint. Suppose prior knowledge suggests that interactions up to order 3 involving a certain component are likely important, whereas even two-way interactions involving some other components are negligible. One can still use a carefully chosen FFD [107].

One setting in which factorial designs are not well suited is when the main effects of all the individual components are weak, but there are some high-order interactions in the data-generating model that produce a strong effect on the outcome (i.e., a setting where the *Hierarchical Ordering Principle* is violated). Another important caveat regarding the use of factorial designs for developing multicomponent interventions is the presence of nested components (e.g. levels of component B are nested within the levels of component A). Generalization of the usual factorial designs called *nested factorials* [85, 3] can incorporate nested components. Analysis of such designs can employ mixed-effects models [82]. A somewhat similar issue is when some intervention components are applied most naturally in a grouped setting. For example, some intervention components are provided to all patients at a clinic [30] or to all children in a classroom or school [37]. Development of an experimental framework tailored to such settings is an avenue for future research. To conclude, FFDs provide

a powerful tool for conducting screening studies to aid in the development of multi-component interventions.

3.6 Appendix A: Defining Relation and Resolution of an FFD

The *defining relation* of an FFD specifies the aliasing. Suppose a study involving five components, say $A_1, \ldots, A_5$, is restricted to 16 cells (as in the *Guide to Decide* study described in [62]). Then a $\frac{1}{2}$ fraction of the 2^5 full factorial design should be used. With 16 cells, one can construct a full factorial with 4 components, say with $A_1, \ldots, A_4$. The strategy is to alias the fifth component, say A_5, with the 4-way interaction $A_1A_2A_3A_4$. This means, the column (in the design matrix) of A_5 is identical to that of the element-wise product of the columns of A_1, A_2, A_3, and A_4, i.e., $A_5 = A_1A_2A_3A_4$. Note that all the elements in any of the columns are either $+1$ or -1. So element-wise product of any column with itself leads to the identity column, say I (with all its entries $+1$). In particular, $A_5A_5 = I$. Multiplying both sides of the equation $A_5 = A_1A_2A_3A_4$ by A_5 gives

$$I = A_1A_2A_3A_4A_5. \tag{3.4}$$

The condition (3.4) completely specifies the aliasing pattern of the 2^{5-1} FFD under consideration, and hence called its *defining relation*. The alias of any factorial effect can be found by multiplying both sides of (3.4) by that effect and then using the facts that $A_jI = A_j$ and $A_jA_j = I$ for all j. The word $A_1A_2A_3A_4A_5$ is called the *defining word*. The length (i.e., number of elements) of the *defining word* is called the *resolution* of the design. So the design specified by (3.4) is a resolution V design. In a resolution V design, the main effects are aliased with 4-way interactions, and the 2-way interactions are aliased with 3-way interactions. In a setting where the third-

or higher-order interactions are negligible, resolution V FFDs are almost as good as the full factorials in that the main effects and 2-way interactions are estimable without bias.

However due to cost and feasibility constraints, one often has to use smaller (than $\frac{1}{2}$) fraction of full factorial designs, leading to lower resolution. The *Project Quit* study described before used a resolution IV FFD. In the following, we illustrate resolution IV designs with an example. Suppose there are 6 components, say $A_1, \ldots, A_6$ in a study that is restricted to 16 cells (as in *Project Quit*). This means constructing a $\frac{1}{4}$ fraction of the 2^6 (= 64 cells) full factorial design. With 16 cells, one can construct a full factorial with 4 components, say with $A_1, \ldots, A_4$. Now, the strategy is to make the columns of the remaining two components A_5 and A_6 identical to some higher-order interactions. One such choice is to set $A_5 = A_1A_3A_4$ and $A_6 = A_2A_3A_4$. Using the same rules as before, one gets $I = A_1A_3A_4A_5$ from the first aliasing relation, and $I = A_2A_3A_4A_6$ from the second aliasing relation. Multiplying these two, a third equation $I = A_1A_2A_5A_6$ follows. Thus the defining relation of this FFD is

$$I = A_1A_3A_4A_5 = A_2A_3A_4A_6 = A_1A_2A_5A_6. \tag{3.5}$$

By definition, (3.5) is a resolution IV design, since the length of each defining word is 4. In a resolution IV design, the main effects are aliased with 3-way or higher-order interactions, but the 2-way interactions are aliased with other 2-ways.

3.7 Appendix B: Relative Efficiency of The Two Ways of Forming The Test Statistic

Here we show that if there are k (≥ 2) components in a factorial experiment, and there may be a two-way but no higher-order interaction in the true data-generating

model, then the relative efficiency of the two ways of forming the test statistic (measured by η) increases with k.

Note that in 0/1 coding, regardless of the total number of components (k), β_1, the coefficient of A_1, is a simple effect given by $\mu_{(1,0,\ldots,0)} - \mu_{(0,\ldots,0)}$; the corresponding estimator is $\hat{\beta}_1 = \bar{Y}_{(1,0,\ldots,0)} - \bar{Y}_{(0,\ldots,0)}$, where $\bar{Y}_{(1,0,\ldots,0)}$ is the sample mean of Y in the $(1, 0, \ldots, 0)$ cell, and so on. It follows that $E(\hat{\beta}_1) = b_1$, and $Var(\hat{\beta}_1) = 2\sigma^2/r$ (since $\hat{\beta}_1$ is a comparison of two cells each of size r). So when using simple effects to form the test statistic, the signal-to-noise ratio governing the power to detect A_1 in the 2^k design is same as that in a 2×2 design considered before, e.g.,

$$SNR_{\text{simple}} = \frac{|\mathrm{E}(\hat{\beta}_1)|}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} = \frac{|b_1|\sqrt{r}}{\sqrt{2}\sigma}.$$

In $-1/1$ coding,

$$\beta_1 = \frac{1}{2} \times (\text{the main effect of } A_1) = \frac{1}{2} \times (\text{the average of } 2^{k-1} \text{ simple effects}),$$

which is estimated by its sample version $\hat{\beta}_1$ where all the μ's in the expression of simple effects are replaced by the corresponding $\bar{Y}$'s. In case there is a two-way interaction (b_{12}) but no higher-order interaction in the true data-generating model, $E(\hat{\beta}_1) = \beta_1 = (\frac{b_1}{2} + \frac{b_{12}}{4})$,

$$\begin{aligned}
Var(\hat{\beta}_1) &= \frac{1}{4} \times \frac{1}{2^{k-1}} \times (\text{variance of an estimated simple effect}) \\
&= \frac{1}{4} \times \frac{1}{2^{k-1}} \times \frac{2\sigma^2}{r} = \frac{\sigma^2}{2^k r}, \\
\text{and} \quad SNR_{\text{main}} &= \frac{|\mathrm{E}(\hat{\beta}_1)|}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} = 2^{(k/2-1)} \left| b_1 + \frac{b_{12}}{2} \right| \frac{\sqrt{r}}{\sigma}. \\
\text{Thus,} \quad \eta &= \frac{SNR_{\text{main}}}{SNR_{\text{simple}}} = 2^{(k-1)/2} \left| 1 + \frac{b_{12}}{2b_1} \right|, \quad \text{an increasing function of } k.
\end{aligned}$$

CHAPTER IV

Comparison of the MOST Approach and a Single Randomized Trial for Developing Multicomponent Interventions

This chapter provides a head-to-head comparison between two competing approaches to developing multicomponent interventions: (a) the *classical* approach consisting of a standard two-group randomized trial followed by post hoc analyses, and (b) the MOST approach. Here we present results from a simulation study in which the classical and the MOST approaches were applied to the same randomly generated data. As we will see, the MOST approach resulted in better mean intervention outcomes under medium or large effect size, whereas the classical approach resulted in better mean intervention outcomes when the effect size was small.

4.1 Introduction

In this chapter, we describe a simulation study to contrast and explicate the relative advantages and disadvantages of two different approaches to empirically building and evaluating multicomponent interventions. The more typically used approach, labeled here the *classical* approach (sometimes called the *treatment package strategy*), consists of constructing a likely best intervention a priori, evaluating the interven-

tion in a standard randomized controlled trial, and following it up with observational analyses. On the other hand, the emergent MOST approach [25, 62] discussed in Chapter I involves programmatic phases of empirical research and discovery aimed at identifying individual component effects and the best combination of components and levels. The MOST approach makes use of fractional factorial designs (FFD) discussed in Chapter III.

This chapter describes a simulation study that addresses the following questions: (1) Which approach, the classical or the MOST, is better at identifying (a) more efficacious interventions? (b) the correct set of intervention components and levels? (c) the best setting of a component with several possible settings? (d) the active components that should be included? (e) the inactive components that should be excluded? (2) What is the impact of overall intervention effect size on the absolute and relative performance of the two approaches? We also briefly summarize the results of additional simulations performed to assess the generalizability of the results.

4.2 Methods

4.2.1 Overview of the Simulation

In this simulation the behavioral scientist intends to build and evaluate a multicomponent intervention. Based on existing literature, prior study results and clinical experience, the scientist has identified five intervention components, denoted $A_1 - A_5$, each of which is hypothesized to have a positive effect on an outcome variable Y. Components $A_2 - A_5$ can be either included in the intervention or not included, thus they can assume only two levels. A_1 can assume three levels: low, medium, or high. In building and evaluating the behavioral intervention the scientist has access to $N = 1200$ subjects. We chose $N = 1200$ because it is not uncommon for behavioral

intervention trials to have sample sizes at least this large; examples include [88], which had $N = 1848$ subjects; [89], which had $N = 3971$ subjects; and [77], which had $N = 1421$ subjects. In appendix C, we summarize the results of additional simulations using smaller and larger sample sizes.

Data sets were generated using a procedure (described below) designed to reflect some of the complexity that can occur in real intervention studies. Both the classical approach and the MOST approach were separately applied to each generated data set. The goal of each approach was to arrive at the most efficacious intervention, expressed in terms of an outcome variable Y. The classical approach consisted of selecting components and dosages a priori and performing a two-group randomized trial using all available subjects. This was followed by post-hoc analyses. By contrast, the MOST approach began with an initial screening experiment for preliminary selection of components, based on a portion of the sample. This was followed by a set of refining experiments to finalize selection of components and dosages, based on the remaining portion of the sample.

4.2.2 Data Generation Model

The data generation model used in this simulation study was inspired by the conceptual model used in a large behavioral intervention trial called *Fast Track* [27]. This data generation model was designed to be only partially consistent with the behavioral scientist's hypotheses described above, in order to mimic the commonly occurring real-life situation in which some of an investigator's hypotheses are true and some are false. Although the investigator hypothesized that all five components would have a positive effect on the outcome, in the data generation model the only active components were A_1, A_2, and A_4. In addition, the relation between A_1 and Y

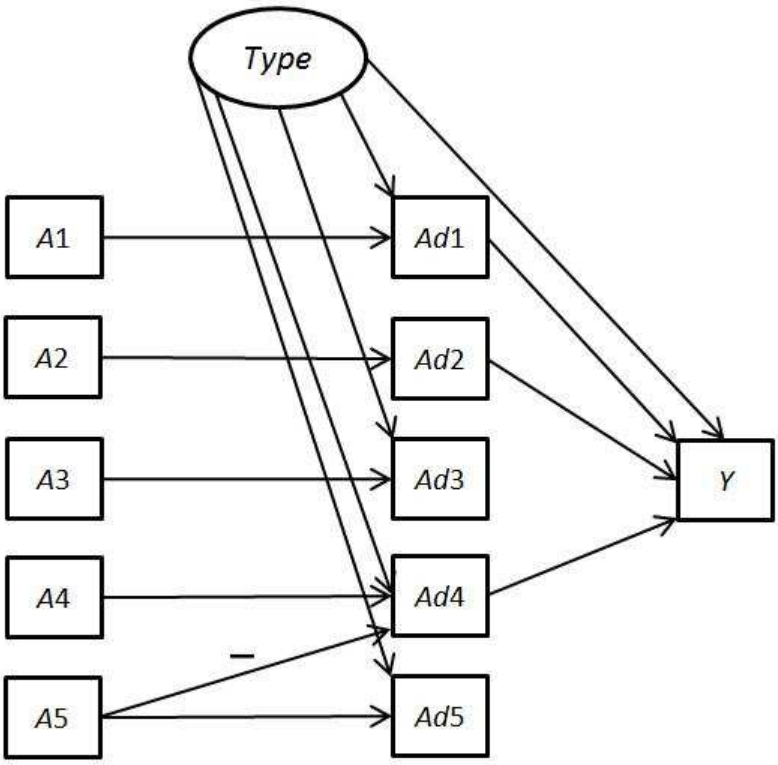

Figure 4.1: Data generation model for simulation.

was curvilinear such that the medium level of A_1 was associated with higher values of Y than other levels of A_1 (and the high level of A_1 was associated with higher values of Y than the low level of A_1). Thus the optimal configuration of intervention components is A_1 included in the intervention and set to the medium level; A_2 and A_4 included; and A_3 and A_5 not included.

Additional complexity was introduced in the data generation model in three different ways to reflect circumstances that frequently occur in real-world intervention settings. First, to represent the amount of each component actually received by (rather than assigned to) participants, adherence variables $Ad_1 - Ad_5$ were modeled for each of the five components $A_1 - A_5$. Adherence was modeled as 100% for A_2 (i.e., $Ad_2 = A_2$) and partial for the remaining components ($0 \leq Ad \leq A$); see Appendix

A for details. Second, to mimic the confounding that can result when post-hoc analyses use non-randomized comparisons, unknown participant characteristics that can affect both adherence and the outcome were included. In a real-life setting, there are likely to be many such confounding variables. For simplicity, they were modeled here by a single unobserved binary variable called *Type*, with $Type = 1$ representing participants likely to register a higher value of Y, and $Type = 0$ representing participants likely to register a lower value of Y. In addition to its relation with Y, *Type* is positively associated with the level of adherence (except Ad_2 which is always 100%) so that participants are more likely to adhere and hence receive more treatment if $Type = 1$. Thus, *Type* causes a spurious positive correlation between the levels of adherence (except Ad_2) and Y, which in turn makes the estimates of component effects based on the post-hoc analyses positively biased. Third, when participants are offered multiple behavioral intervention components of varying attractiveness, some may adhere closely to the more attractive components and reduce their adherence to the others. This has a deleterious effect if some of the less attractive components are more efficacious. To mimic this, a negative interaction was modeled between A_4 and A_5, such that A_5 induced a reduced adherence to A_4. This means that all else being equal, an intervention that included both A_4 and A_5 is less efficacious than one that included A_4 without A_5. This phenomenon is called *subadditivity*.

Figure 1 is a pictorial representation of the data generation model, details of which are provided in Appendix A. Figure 1 is a directed acyclic graph [66]; the presence of an arrow from one variable to another indicates that the former variable may have a causal effect on the latter variable. A square represents an observed variable, and a circle represents an unknown, and hence unobserved, variable. The absence of an arrow indicates conditional independence; for example, given the variable Ad_1, Y

is independent of A_1. In Figure 1 all relations have a positive (if any) dependency except the A_5 to Ad_4 relation, which is labeled with a minus sign. To maintain simplicity and clarity of exposition no other population heterogeneity was built into the simulation. Thus in the following analyses it would not be useful to control for observed participant characteristics or other observed pretreatment variables. Averaging over the distribution of *Type* and $Ad_1 - Ad_5$ produces the marginal linear model

$$E[Y|A_1, \ldots, A5] = c_0 + c_1 A_1 + c_{11} A_1^2 + c_2 A_2 + c_4 A_4 + c_{45} A_4 A_5 \tag{4.1}$$

Furthermore the variance of Y given $A_1, \ldots, A_5$ is a function of the components $A_1, \ldots, A_5$, that is, the variance is nonconstant (see Appendix A).

4.2.3 Experimental conditions

In the simulation there were three effect size conditions for the interventions, corresponding to Cohen's [21] benchmark values for standardized effect sizes of small ($d = 0.2$), medium ($d = 0.5$) and large ($d = 0.8$). Effect sizes were defined in terms of the ideal intervention as a whole, in other words, for the two-group comparison of the best treatment combination (A_1 set to medium, A_2 and A_4 included, A_3 and A_5 not included) versus a control group. Active main effects were roughly equal in magnitude, and the effect corresponding to the active interaction ($A_4 A_5$) was roughly half the size of the main effects. All other intervention component main effects and interactions were set to 0. Across all three effect size conditions, the effect of *Type* was set equal to $d = 0.9$. For each of the three effect size conditions, 1000 simulated data sets of $N = 1200$ random experimental subjects were generated. Each generated data set was used twice: once for the classical approach, and once for the

MOST approach. All the results presented in Tables 4.1-4.3 are averages based on the 1000 simulated data sets.

4.2.4 Operationalization of the classical and the MOST approaches

This section contains a brief overview of the operationalizations of the classical and the MOST approaches. A detailed description of the MOST approach can be found in Appendix B.

The classical approach

The classical approach employed all $N = 1200$ experimental subjects in a single randomized trial of the multicomponent treatment vs. control, followed by post hoc analysis. The treatment group was given an intervention consisting of A_1 set to "high" and all of the other components included; the control group was given an intervention with A_1 set to "low" and none of the other components included. Because only homogeneous subpopulations were considered in this simulation study (see the data generation model above), there were no pretreatment variables to be controlled. As is traditional, a two-group comparison was performed for the overall efficacy of the intervention. However, regardless of the outcome of this comparison, decisions about whether individual components should be retained in the intervention were based on post-hoc dose-response analyses (with the levels of adherence $Ad_1, \ldots, Ad_5$ as doses) on the treatment group of subjects as follows:

Step 1: Identify components with sufficient variation in dose to enable dose-response analyses. Received dose (adherence) could vary between 0 and 2 for A_1 and between 0 and 1 for $A_3 - A_5$ (adherence was always 100 percent for A_2, so there

was no variation). Any components for which naturally occurring variation in dose was greater than an arbitrary threshold of 0.01 were considered to have sufficient variation to enable dose-response analyses. Any components with variation in dose less than 0.01 could not be examined further, and were automatically included in the final intervention.

Step 2: Multiple regression. The outcome Y was regressed on the following variables: doses of all components with sufficient variation in dose; two-way interactions between them; and, if Ad_1 was included in the regression, a quadratic term for Ad_1 (an implicit assumption here is that the scientist knows that A_1 has more than two levels).

Step 3: Select components and levels. The estimated regression function was evaluated at each combination of levels of the components that had sufficient variation in dose (by plugging in possible values of A's in place of Ad's, e.g., 0, 1, or 2 for Ad_1, and 0 or 1 for $Ad_3 - Ad_5$). The level combination that produced the largest predicted value of Y was identified.

Step 4: Final intervention. The final intervention identified by the classical approach consisted of (a) the low-variation components identified in Step 1, each set to 1 (2 in the case of A_1), plus (b) the configuration of components and levels identified in Step 3.

The MOST approach

The MOST approach used the same $N = 1200$ subjects as the classical approach,

but employed $N = 800$ subjects in an initial screening phase and reserved $N = 400$ for a subsequent refining phase. In the screening phase, a factorial experiment involving all five components was conducted, with only the low and the high levels of A_1 included. To conserve resources, a 16-condition balanced fractional factorial design was used instead of a 32-condition complete factorial; see Appendix B for a technical discussion about the particular choice of fraction and the rationale behind it. This experiment was used to identify significant main effects and 2-way interactions. Intervention components were selected based on the results of the screening phase using the following decision rules. First, any component with a significant main effect and not involved in a significant interaction was selected for inclusion in the intervention. A main effect was deemed significant if it possessed one of the three largest positive t-statistics or if the associated t-test was significant at the 0.10 level and positive. The decision rule to take the three largest was arbitrary to an extent; below we summarize the results of additional simulations that varied this decision rule. Interactions were deemed significant if the associated t-test was significant at the 0.10 level. Next, any components involved in significant two-component interactions were examined further. The combination of the two components that produced the highest marginal cell mean on Y was selected for inclusion. This procedure is described in more detail in Appendix B.

In this study the purpose of the refining phase was to determine the optimal value of A_1. Therefore, if A_1 and all its interactions were insignificant, the refining phase was not conducted. Otherwise, additional experimentation to revise the selected level for A_1 was conducted as follows: (a) If the main effect of A_1 was significant but no interactions were significant, the refining experiment was a two-group com-

parison of level 2 of A_1 against level 1 of A_1. In this experiment the remaining components were set at the levels indicated by the screening phase. The results of this experiment yielded the best level for A_1. (b) If there were one or two significant interactions involving A_1, a factorial experiment was conducted crossing A_1 with the components involved in the interactions, with the remaining components set to the levels indicated by the screening phase. These results yielded the best levels for A_1 and for the components that interacted with A_1. More detail appears in Appendix B.

Evaluation of outcomes of each approach

Because in this simulation the true data generation model is known, it is possible to use this model to evaluate the performance of the classical and the MOST approaches. After the final intervention was determined using either the classical or the MOST approach, the data generation model was used to compute the expectation of the distribution of Y that would be obtained if the intervention were applied to all subjects in the population. These expectations, $E(Y)_{classical}$ and $E(Y)_{MOST}$, were the outcome variables used to evaluate the performance of each approach. In real life this step would instead consist of conducting a large confirmatory randomized trial comparing the final intervention to an appropriate control group. This is called the *confirming* phase of the MOST approach [25, 62].

4.3 Results

As was described above, the classical approach and the MOST approach each identifies a final multicomponent intervention for every simulated data set. The two final multicomponent interventions are then evaluated using the known data

generation model. All the results presented in Tables 4.1-4.3 are averaged over the 1000 simulated data sets.

Table 4.1: Mean Intervention Outcome under Classical and MOST Approaches (averaged over 1000 simulated data sets)

Effect Size	$E(Y)_{classical}$ (standard error)	$E(Y)_{MOST}$ (standard error)	Difference (standard error)	Maximum Possible $E(Y)$
Small	1.72 (0.00)	1.69 (0.01)	0.03 (0.01)	1.99
Medium	2.35 (0.01)	2.58 (0.01)	-0.23 (0.02)	2.99
Large	3.01 (0.02)	3.75 (0.01)	-0.74 (0.02)	4.00

Table 4.1 shows the mean outcome of the classical and MOST approaches, the mean difference between them, and standard errors. For reference, the maximum possible mean outcome value is included. Table 4.1 shows that in the small effect size condition $E(Y)_{classical}$ was approximately two percent larger than $E(Y)_{MOST}$, indicating that in this condition the average intervention outcome was slightly better for the classical approach. In the medium and large effect size conditions the average intervention outcome was about 10 and 25 percent larger, respectively, for the MOST approach. The difference between the classical and MOST approaches is significant at the 0.05 level in every condition.

Table 4.2: Comparison of Classical and MOST Approaches on E(Y) (Percentage of Data Sets)

Effect Size	$E(Y)_{classical}$ Higher	$E(Y)_{MOST}$ Higher	Neither Higher (tied)
Small	54.2	40.8	5.0
Medium	32.3	62.4	5.3
Large	14.9	75.7	9.4

Table 4.2 shows the percent of data sets in which each approach "won" by identifying an intervention that yielded a larger value of the outcome $E(Y)$. In the small effect size condition the classical approach was about 1.3 times more likely than the MOST approach to identify an intervention that yielded a larger $E(Y)$. In the medium and large effect size conditions the effect was reversed, with the MOST ap-

proach about 1.9 and 5.1 times more likely, respectively, to identify an intervention that yielded a larger $E(Y)$.

Table 4.3: Accuracy of Component Selection under Classical and MOST Approaches (Percentage of Data Sets)

Effect Size	Classical	MOST
Correct Combination of Components/Levels Identified		
Small	1.9	7.5
Medium	3.7	24.3
Large	5.5	52.0
All Active Components Identified		
Small	48.5	14.5
Medium	48.4	37.3
Large	48.2	73.5
All Inactive Components Identified		
Small	20.0	45.7
Medium	19.5	61.0
Large	19.2	68.5

Table 4.3 depicts the accuracy with which each approach selected intervention components and levels for inclusion in the intervention or identified components for exclusion. The first section of the table shows the percent of data sets in which the correct configuration of components and levels was identified. As expected, this number increased for both approaches as effect size increased. In every condition the MOST approach was much more likely to identify the correct configuration. One reason for the better performance of the MOST approach is that it identified the medium level of A_1 as optimal more frequently than the classical approach (in 61.3 vs. 11.8 percent of data sets in the small effect size condition, 90.7 vs. 31.0 percent in the medium effect size condition, and 98.3 vs. 40.6 percent in the large effect size condition). Another reason is that the MOST approach included the component A_5 (which, as described above, produced a subadditive effect in presence of the active component A_4) in the intervention much less frequently (in 25.4 vs. 57.6 percent of data sets in the small effect size condition, 11.5 vs. 57.2 percent in the medium effect

size condition, and 6.5 vs. 56.9 percent in the large effect size condition).

The second section of Table 4.3 shows the percentage of data sets in which all active components were correctly selected, irrespective of whether inactive components were mistakenly selected, and irrespective of the selected level of A_1. The classical approach outperformed the MOST approach on this criterion for the small and medium effect size conditions. For the MOST approach the performance improved dramatically (from 14.5 to 73.5 percent) as the effect size increased. However, the performance of the classical approach was fairly constant (ranging from 48.2 to 48.5 percent) across the effect size conditions.

The third section of Table 4.3 shows the percentage of data sets in which all inactive components were correctly excluded, irrespective of whether some active intervention components were incorrectly excluded. Across all three effect size conditions the performance of the MOST approach was better than the classical approach. Here too performance improved as effect size increased for the MOST approach, but not for the classical approach.

Other sample sizes and numbers of main effects retained

We conducted some additional simulations in order to investigate whether the results reported here held across variation along two dimensions. One was sample size. The other was the decision rule used in the MOST approach for selecting intervention components for inclusion based on main effects estimates. In a series of nine simulations we investigated three different sample sizes, $N = 600$, $N = 1200$, and $N = 2500$; and three different decision rules: retention of the intervention components corresponding to the largest two, three, and four main effects. The overall pattern of results was very consistent. In general the classical approach tended to

produce a larger $E(Y)$ than the MOST approach in conditions involving both a small effect size and a small sample size. The MOST approach tended to produce a larger $E(Y)$ than the classical approach in the medium and large effect size conditions, even in the small sample size condition. The MOST approach tended to produce larger $E(Y)$ than the classical approach when the decision rule called for retaining a larger number of main effects; in the conditions in which the four largest main effects were retained the MOST approach consistently produced the larger $E(Y)$, even in the conditions involving both a small effect size and a small sample size. More details can be found in Appendix C.

4.4 Discussion

The simulation reported here compared one possible operationalization of the MOST approach to one possible operationalization of the classical approach. Which approach performed best depended upon which criterion was used to evaluate the approaches and also upon intervention effect size.

When the two approaches were evaluated in terms of overall intervention outcome, the classical approach performed better than the MOST approach when the intervention effect size was small, and the MOST approach performed better than the classical approach when the intervention effect size was medium or large. The MOST approach suffered somewhat from a lack of power in the small effect size condition. One reason why the classical approach tended to be outperformed by the MOST approach in the medium and large effect size conditions is confounding by the unknown participant characteristic *Type*. *Type* introduced a positive bias that had an impact on the results of the classical approach primarily in two areas. First, this positive

bias made the high level of A_1 look better than the medium level in the post-hoc dose-response analysis. Second, the positive bias also masked the subadditive effect of A_5 on A_4 in the post-hoc analysis, sometimes leading to the incorrect inclusion of the component A_5. *Type* had little or no impact on the results of the MOST approach because this approach depended primarily on estimates of main effects and interactions based on data from randomized experiments, which are much less likely to be biased by confounding than are post hoc non-experimental analyses [83].

When success at identifying the best combination of components and levels was the criterion, the MOST approach was the better of the two across all effect sizes. This is directly due to the greater impact of confounding on the classical approach as compared to the MOST approach. For example, confounding by *Type* made the classical approach more likely to lead to choose an incorrect level of A_1, as mentioned above, even though a quadratic term was appropriately included in regression analyses.

When the two approaches were evaluated in terms of successfully including all of the active components, the classical approach performed better than the MOST approach in the small and medium effect size conditions, and the MOST approach performed better when effect sizes were large. The MOST approach detected active components at a higher rate as effect size increased, due to the corresponding increase in power. By contrast, the classical approach detected active components at a relatively constant rate across increasing effect sizes. The primary reason for this is that in our operationalization of the classical approach the components with low variability in adherence were automatically included, irrespective of effect size. For example, the active component A_2 was always included because the received dose, Ad_2, was always equal to the assigned value of A_2 (100% adherence).

The MOST approach outperformed the classical approach in all effect size conditions when the criterion was successfully identifying inactive or potentially counterproductive components that should be eliminated from the intervention. Again, this is attributable to the differential impact of confounding. Because *Type* is positively associated with both Y and Ad's, it induced a positive bias in the non-experimental analyses that led the classical approach to a preference for including components over excluding them.

Choosing an approach to intervention building

Our results suggest that when medium or large intervention effect sizes are anticipated, the use of the MOST approach is likely to result in identifying a more potent intervention than the classical approach. When a small intervention effect is anticipated, the choice is less clear. Multicomponent interventions with small overall effect sizes may be made up of either (a) mostly inactive components with one or two components with relatively large effect sizes, or (b) fairly equally efficacious but weak components, which together produce a detectable aggregate effect even though no individual component has a detectable effect. In situation (a), the MOST approach may be helpful in identifying the inactive components. In situation (b), in order to perform well the MOST approach would need to be powered to detect the weak individual component effects. Here the classical approach may be a better choice.

One drawback of the classical approach is that all subjects in the treatment arm receive all the components, and so the main effects of all the components and their interactions of every order are confounded (aliased). In contrast, in the MOST approach if a fractional factorial design is used, the main effects and two-way interactions are confounded with only higher order interactions deemed negligible in

size, and hence do not affect the results. If a full factorial design is used there is no confounding of main effects and interactions.

It is possible that some other variant of the classical method, e.g., dismantling experiments [103], would have performed better than the approach used in the simulation reported here. However, as long as post hoc analyses on non-randomized data (e.g., adherence) are used, the performance of any version of the classical method will depend on the degree of confounding present in the data. In situations in which the degree of confounding is very low or nonexistent, the version of the classical approach we have used and other reasonable variants would probably perform as well as the MOST approach. Of course, in most cases the degree and nature of confounding is not under the investigator's control and may be difficult to anticipate. Because the MOST approach is entirely based on randomization, it is much less vulnerable to confounding.

Although the MOST approach was better overall at identifying the best configuration of components and levels, the success rate ranged from a high of 52 percent to a low of about 7.5 percent. Thus there is plenty of room for improvement, particularly when effect sizes are small. It is possible that an augmented approach or even an entirely different approach could result in a higher success rate. One promising avenue for intervention refinement may lie in exploring ideas from engineering process control, as discussed in [69].

Differences in resource requirements

One question that arises in considering the MOST approach is whether the additional experimentation required by this approach necessarily demands an increase in cost over the classical approach. The MOST approach calls for a design that can

isolate the effects of individual components. In the screening phase this will usually be some variation of a factorial design, requiring implementation of numerous conditions, each of which represents a different version of the intervention. For example, a full factorial design involving k two-level components requires implementation of 2^k treatment conditions, which may be costly. By contrast, irrespective of the number of components studied the classical approach typically requires implementation of only two conditions, a treatment and a control.

Two important costs are experimental subjects and implementation of experimental conditions. In this simulation the MOST approach used exactly the same number of experimental subjects as the classical approach, suggesting that it is no more demanding with respect to sample size. When a factorial design is used, given a fixed number of subjects an investigator may test as many components as desired – the power to detect every main effect in this way is about the same as testing that component in a single-component two-group study with the same sample size. This means that factorial experiments make very efficient use of experimental subjects. Power and sample size considerations (e.g., for testing main effects) in a factorial setting can be found in [16, 42, 53, 17]; see also Chapter III.

However, even when in the MOST approach the same number of subjects is used as in a comparable classical approach, there may be additional costs associated with implementing a wider variety of versions of the intervention and conducting follow-up experiments. It may also take more time to implement the MOST approach, and may require more training of intervention delivery staff. Although these logistics and costs associated with the MOST approach are a serious consideration, highly efficient fractional factorial designs offer a way to keep the number of experimental conditions manageable. Some assumptions about higher-order interactions being negligible are

necessary in order to take advantage of the economy offered by a fractional factorial over a full factorial design. The particular fractional design used in the simulations reported here did not allow estimation of 3-way or higher order interactions; instead, it required the assumption that they are negligible in size. Many fractional factorial designs are available. If prior knowledge suggests that some 3-way interactions can be important, an investigator can choose a different fractional design that allows estimation of 3-way interactions [107].

Even with a highly efficient design, the investigators in empirical settings sometimes have to make decisions based on results of past studies and auxiliary analysis performed on data from the current study. This is particularly true when numerous interactions are anticipated. An example of such an analysis can be found in [88].

The short-term costs of building and evaluating an intervention must be weighed against long-range costs and benefits. Our results suggest that the MOST approach may help identify more efficient and streamlined interventions by identifying inactive components for elimination. As Allore et al. [1, p. 14] noted, "Since each component of an intervention adds to the overall cost and complexity, being able to directly estimate component effects could greatly enhance efficiency by reducing the number of components introduced into clinical practice". Our results also suggest that under many circumstances the MOST approach may be likely to identify a more efficacious intervention than the classical approach. Thus, in some applications the long-range gains in terms of increased efficiency and public health benefits expected to result from the MOST approach may offset any additional up-front intervention development costs.

4.4.1 Limitations

This simulation was designed to take an initial look at the question of whether the MOST approach is a reasonable way to build interventions. It involved only a very small set of conditions out of the infinite number of possibilities that can occur in practice. There are a number of potentially important factors that were not varied in the simulation. A few of these are: the underlying structural model, which could be varied to include features such as more 2-way interactions, higher-order interactions, and the presence of mediating variables; the degree of confounding, here reflected by the variable *Type*; the number of components under consideration; the number of active vs. inactive components; other effect sizes besides the three used here; the impact of measurement noise on the outcome variable; the effect of complex data structures such as nesting (e.g. individuals within classrooms; patients within clinics); incorporating cost and burden in decisions about which components and levels should make up an intervention; and the operationalization of the classical approach used. Many other additional factors could be considered. Despite the limitations of this study and the need for additional research, we believe that the results of the simulation show clearly that the MOST approach is a promising alternative.

4.4.2 Conclusions

The classical approach is currently the most well-established approach to empirical development of behavioral interventions. However, an emergent strategy, labeled here the MOST approach, provides a systematic way of making evidence-based decisions about which components and which levels of the components should comprise an intervention. Comparison of the two approaches in real-world empirical settings

is impractical. In the present chapter a simulation was presented that provides this comparison by modeling a plausible empirical scenario. The results suggested that the MOST approach merits serious consideration, because it has the potential to help intervention scientists to build more potent behavioral interventions. Possible exceptions to this are interventions with a small overall effect size, particularly those that are the cumulative effect of many weak components. More research is needed on methods to identify the optimal intervention, and thereby increase public health benefits.

4.5 Appendix A: Data Generating Model

The data generating model is described below in terms of equations involving the intervention components ($A_1 - A_5$), the measures of adherence ($Ad_1 - Ad_5$), an unknown confounder *Type* (T), and the outcome (Y). In this model, $A_2 - A_5$ can take two values: 0 or 1 (absent or present); while A_1 can be 0, 1, or 2 (low, medium, or high). Note that subjects may receive a different dose of a component than that assigned. Measures of adherence (Ad's) simply represent these doses. A multiplicative model is used below to describe the relation between A's and Ad's. The confounder *Type* follows a Bernoulli (1/2) distribution.

$\underline{A \rightarrow Ad}$:

$$
\begin{aligned}
Ad_1 &= (\eta_{10} + \eta_{11}\, T + e_1) \cdot A_1 \\
Ad_2 &= A_2 \\
Ad_3 &= (\eta_{30} + \eta_{31}\, T + e_3) \cdot A_3 \\
Ad_4 &= (\eta_{40} + \eta_{41}\, T + \eta_{42} A_5 + e_4) \cdot A_4
\end{aligned}
$$

$$Ad_5 \quad = \quad (\eta_{50} + \eta_{51}\, T + e_5) \cdot A_5$$

where each of e_1, e_3, e_4, and e_5 follows a normal distribution $N(0, \sigma_e^2)$, $\sigma_e = 0.1$. Note that there is no non-adherence to component A_2. The right hand side of the equations in the above display are truncated such that $Ad_j \in [0, A_j], \forall j$. The subsequent equations are only approximate due to this truncation.

$\underline{Ad \to Y}$:

$Y = \beta_1\, T + \beta_2\, Ad_1 + \beta_3\, Ad_1^2 + \beta_4\, Ad_2 + \beta_5\, Ad_4 + \epsilon_Y; \quad \epsilon_Y \sim N(0, 3).$

Marginal Form of Y, averaged over Ad's:

Averaging over Ad's, we get

$$\begin{aligned} Y \;=\; & \beta_1\, T + \beta_2\, (\eta_{10} + \eta_{11}\, T)\, A_1 + \beta_3 \left((\eta_{10} + \eta_{11}\, T)^2 + e_1^2 \right) A_1^2 + \beta_4\, A_2 \\ & + \beta_5\, (\eta_{40} + \eta_{41}\, T)\, A_4 + \beta_5\, \eta_{42}\, A_4\, A_5 \\ & + \left(\epsilon_Y + e_1\, \beta_2\, A_1 + 2 e_1\, \beta_3\, (\eta_{10} + \eta_{11}\, T)\, A_1^2 + e_4\, \beta_5\, A_4 \right) \end{aligned} \tag{4.2}$$

Let e_T denote the sum of the 4 terms in the last row of the above display. Note that the term e_T has zero mean but heteroscedastic variance because some of the e_j's occur in products with the components. Because of zero mean, e_T functions like an error term. The generated Y will have a mean of the form

$$\begin{aligned} E[Y|A_1, \ldots, A_5] \;=\; & \frac{1}{2}\beta_1 + \beta_2\, (\eta_{10} + \frac{1}{2}\eta_{11})\, A_1 + \beta_3\, (\eta_{10}^2 + \frac{1}{2}\eta_{11}^2 + \eta_{10}\, \eta_{11} + \sigma_e^2) A_1^2 \\ & + \beta_4\, A_2 + \beta_5\, (\eta_{40} + \frac{1}{2}\eta_{41})\, A_4 + \beta_5\, \eta_{42}\, A_4\, A_5 \\ \;=\; & c_0 + c_1\, A_1 + c_{11}\, A_1^2 + c_2\, A_2 + c_4\, A_4 + c_{45}\, A_4\, A_5 \end{aligned} \tag{4.3}$$

where each c_j is a function of (η, β, σ_e). In the simulations, we set these parameter values to ensure certain effect sizes as defined in the following section. Equation (4.3)

above corresponds to equation (4.1) appearing in the main text.

Standardized Effect Size

In our simulations, we set the parameter values so that the standardized effect size (Cohen's d) for the two-group comparison of the best treatment combination ($A_1 = A_2 = A_4 = 1$, $A_3 = A_5 = 0$) vs. the control where A_1 is set to its low level and all other components are absent (i.e., $A_i = 0, \forall i$) enjoys Cohen's benchmark values (small=0.2, medium=0.5, large=0.8). Cohen's d in this case is explicitly defined as:

$$\begin{aligned} d &= \frac{E\Big(Y|A_1 = A_2 = A_4 = 1; A_3 = A_5 = 0\Big) - E\Big(Y|A_i = 0, \forall i\Big)}{\sqrt{\frac{1}{2}\Big[\mathrm{Var}\Big(Y|A_1 = A_2 = A_4 = 1; A_3 = A_5 = 0\Big) + \mathrm{Var}\Big(Y|A_i = 0, \forall i\Big)\Big]}} \\ &= \frac{c_1 + c_{11} + c_2 + c_4}{f(\eta, \beta, \sigma_e)}, \end{aligned} \tag{4.4}$$

where f is some function of η, β, σ_e as can be seen from (4.2). The numerator follows from (4.3).

Parameter Values for Specific Standardized Effect Sizes

From now on we denote the parameter values used in a given simulation with a superscript 0. The true parameter values η^0 and β^0 are chosen so that the following conditions are satisfied:

1. The standardized effect size d as defined above attains Cohen's benchmark values (small=0.2, medium=0.5, large=0.8).

2. The active main effects (considered in the screening phase) are roughly equal in

magnitude, while the active interaction is half the size of active main effects:

$$2c_1 + 4c_{11} = c_2 = c_4 = c, \quad \text{say} \tag{4.5}$$
$$c_{45} = -\frac{c}{2}$$

3. The middle level of A_1 (i.e., $A_1 = 1$) is best, and the main effect of A_1 between levels 1 and 2 (as considered in refining phase) is also equal to the main effects considered in screening phase, i.e.,

$$(c_1 + c_{11}) - (2c_1 + 4c_{11}) = -c_1 - 3c_{11} = c \tag{4.6}$$

4. The level of confounding, as quantified by $\beta_1\eta_{11}$ $(= \beta_1\eta_{31} = \beta_1\eta_{41} = \beta_1\eta_{51})$, is made equal to c corresponding to the large effect size ($d = 0.8$).

The values of the η^0 stay the same across different effect sizes and are set to:

$$\eta^0_{10} = \eta^0_{30} = \eta^0_{40} = \eta^0_{50} = 0.50;$$
$$\eta^0_{11} = \eta^0_{31} = \eta^0_{41} = \eta^0_{51} = 0.25;\ \eta^0_{42} = -0.3125.$$

If we keep the η's and σ_e fixed, then $f(\eta, \beta, \sigma_e)$ appearing in the denominator of (4.4) can be written as $g(c)$, a function of just c. Also, each β can be expressed as a function of c. From (4.4), (4.5), and (4.6), we get

$$d = \frac{4c}{g(c)} \tag{4.7}$$

For all three values of d (=0.2, 0.5, 0.8), we solve (4.7) for c by recursive method (calculating $g(c)$ by Monte Carlo integration). We get $c = 0.165, 0.415, 0.667$ for small, medium, and large effect size respectively. From the values of c, we can easily obtain the β values.

The β values corresponding to small standardized effect sizes are:

$$\beta_1^0 = 2.6680;\ \beta_2^0 = 0.9240;\ \beta_3^0 = -0.5945;\ \beta_4^0 = 0.1650;\ \beta_5^0 = 0.2640.$$

The β values corresponding to medium standardized effect sizes are:

$$\beta_1^0 = 2.6680;\ \beta_2^0 = 2.3240;\ \beta_3^0 = -1.4953;\ \beta_4^0 = 0.4150;\ \beta_5^0 = 0.6640.$$

The β values corresponding to large standardized effect sizes are:

$$\beta_1^0 = 2.6680;\ \beta_2^0 = 3.7352;\ \beta_3^0 = -2.4033;\ \beta_4^0 = 0.6670;\ \beta_5^0 = 1.0672.$$

4.6 Appendix B: Operationalization of the MOST Approach

As in the classical approach, each scientist following the MOST approach studies all the five components. In the screening phase only the two extreme levels (out of three) of A_1 are considered. We restrict the number of cells to 16 in our simulations, so a 16-cell resolution V balanced fractional factorial design with defining word $I = A_1A_2A_3A_4A_5$ is used (see Appendix A of the Chapter III for a technical discussion on defining word and resolution). The defining word completely specifies the aliasing pattern in the design. The above choice of design was used in a behavioral study on breast cancer prevention [62]. Since this is a resolution V design, the 2-way interactions are not aliased with each other (aliased with three-way interactions, as can be seen from the defining word). In general, the investigators choose the defining word based on prior substantive knowledge concerning the potential strength of higher order interactions relative to the likely noise in the data (e.g. if the size of any three-way interaction is likely small compared to the noise level of the data, one can be more confident that the detected two-interaction effect is due to the two-way interaction and not to a three-way interaction). This is in accordance with the *Hierarchical Ordering Principle* [107, p. 112] which states that, absent strong prior

knowledge, higher order interactions can be expected to be of smaller size than lower order interactions. Note that in the setting described in the main text ($N = 1200$ subjects), only 800 subjects are used by each scientist in the screening phase of the study.

Screening Phase Analysis

The experiment is simulated using the 16-cell balanced fractional factorial design. A standard analysis of variance (ANOVA) is performed on the outcome Y and the five components. In the screening phase, the following steps are followed:

1. A 10% level of significance is used for testing the main effects and two-way interactions (to have greater power).

2. If the no. of significant effects is less than 3, rank-order the absolute values of the t-statistics corresponding to the main effects only (assuming that main effects are more likely to be significant than two-way interactions) and identify the largest 3. Move to the refining phase with the corresponding effects (treating them as significant).

In step 2 above, we chose the number of components to retain (say, k) to be equal to 3 to ensure that at least 50% of the components always pass the screening phase (3 is the smallest integer greater than or equal to 5/2, 5 is the total no. of components). Since in general we expect that only a few components are likely to be active, the choice to carry forward the top 3 components to the refining phase is a reasonable one. By doing so, we are being conservative about the hypothesized effect of the components. This is a tuning parameter of the procedure and can vary from one investigator to another. We have conducted simulations for two other choices of this number (e.g., $k = 2$ and 4). A summary of simulation results across all three choices

of k (e.g., 2, 3, and 4) can be found in Appendix C.

Moving Towards Refining Phase

As mentioned in the main text, a part of the original sample in each simulated data set is reserved for the refining phase. The refining phase may or may not be conducted depending on the results obtained in the screening phase. In general the refining phase is employed in the simulation if (1) the three-level component A_1 is significant in the screening phase, or (2) there is at least one significant interaction involving A_1 (see Algorithm 1 below). The MOST approach, just like the classical approach, assumes the prior knowledge that A_1 is a special component having more than two levels. The refining phase uses multi-group experiments; standard analysis of variance, with 5% level of significance is used. In the setting described in the paper ($N = 1200$ subjects), the remaining 400 subjects are used by each scientist in the refining phase of the study.

Algorithm 1

This algorithm is used in the simulation to determine which components should be retained and which should be rejected, based on the results of the screening phase.
Input: set of components, significant effects (both main and interaction effects), estimated effect sizes, and signs of effects.
Output: best (treatment) combination.
Initialize: *best combination* $= [0, 0, 0, 0, 0]$, a 5-component vector. In the following, we will use the notation *best combination*(i) to denote the i-th element of the vector *best combination*.

1. Go through the set of components ($i = 1 : 5$): If main effect and all interactions of component i are insignificant, set *best combination*$(i) = 0$. If main effect of component i is significant, but none of its interactions are, look at the sign. If sign$(i) = +1$, set *best combination*$(i) = 1$. Else, *best combination*$(i) = 0$.

2. Now for any significant interaction, find $[P_1, P_2]$ = parent components of that interaction.

   (a) If the main effect of the component P_1 (P_2) is insignificant, initialize sign$(P_1) =$ 0 (sign$(P_2) = 0$). If P_1 (P_2) is significant, initialize sign(P_1) (sign(P_2)) to the sign of its main effect (either $+1$ or -1), respectively.

   (b) Define *sign vector* as the vector consisting of sign(P_1), sign(P_2), and sign(interaction).

   (c) If sign$(P_1) = 0$ but sign$(P_2) \neq 0$, set sign$(P_1) =$ sign$(P_2)\times$sign (interaction). Do a similar operation for P_2.

   (d) If both parents are insignificant, i.e., sign$(P_1) =$ sign$(P_2) = 0$, go to step 4.

3. If sign$(P_1)\times$sign$(P_2) =$ sign(interaction), set *best combination*$(P_1) =$ (sign$(P_1)+$ $1)/2$, and *best combination*$(P_2) =$ (sign$(P_2) + 1)/2$.

4. For a significant interaction:

   – if *sign vector*$= [0, 0, 1]$, compare the cell means where both P_1 and P_2 are set to $+1$ (so the interaction P_1P_2 is also set to $+1$) vs. where both P_1 and P_2 are set to -1 (so P_1P_2 is set to $+1$).

   – if *sign vector*$= [0, 0, -1]$, compare the cell means where $P_1 = +1, P_2 = -1$ (so $P_1P_2 = -1$) vs. where $P_1 = -1, P_2 = +1$ (so $P_1P_2 = -1$).

   – otherwise, do cell-mean comparison of the following four cells: $P_1 = +1, P_2 =$ $+1$ (so $P_1P_2 = +1$), $P_1 = -1, P_2 = -1$ (so $P_1P_2 = +1$), $P_1 = -1, P_2 = +1$

(so $P_1P_2 = -1$), $P_1 = +1, P_2 = -1$ (so $P_1P_2 = -1$).

The combination of (P_1, P_2) that gives the highest cell mean is used to determine the best combination. This step implicitly assumes that the four cells do not differ with respect to other active components.

Note that the order in which step 2 considers interactions impacts the results. For example if $\text{sign}(P_1) = 0$, $\text{sign}(P_2) = -1$, $\text{sign}(P_1P_2) = -1$, $\text{sign}(P_3) = +1$, $\text{sign}(P_1P_3) = -1$, then the best combination setting for P_1 will depend on the order in which the interactions are considered. The simulations used the natural ordering, e.g., significant interactions from the ordered list 12, ..., 15, 23, ..., 25, 34, 35, 45 (using the notation ij to denote the interaction A_iA_j).

Refining Phase

1. When A_1 and all its interactions are insignificant, there is no refining phase. The best treatment combination is found by Algorithm 1.

2. When A_1 is significant, but none of its interactions are so, the refining phase uses a two-group follow-up experiment, in which A_1 is varied across the two groups, setting other components at their optimum level (obtained by applying Algorithm 1 on the screening phase results). One group receives the intermediate level of A_1 (not studied in screening phase), the other receives one extreme level depending on the sign of the screening phase estimate of the effect of A_1 (the level is the higher extreme if the sign is a +, lower extreme otherwise). Thus, the best treatment combination is found.

3. If only one interaction involving A_1 is significant, a 6-group follow-up experiment (3 levels of $A_1 \times 2$ levels of the other component forming the interaction), setting

all other components at their optimum levels as found by Algorithm 1, is used.

4. If two interactions involving A_1 are significant, then a 12-group ($3 \times 2^2 = 12$) follow-up experiment is used in the refining phase.

5. For three or more significant interactions involving A_1, conducting follow-up experiment becomes increasingly problematic (constructing many treatment groups), and also results become less reliable (low power for comparing many groups). In our simulations, no refining experiment is conducted in such cases – the best combination is determined by applying Algorithm 1 to the results of screening phase. As mentioned in the previous page, the order in which step 2 of Algorithm 1 considers interactions impacts the results. The simulations used the natural ordering, e.g., significant interactions from the ordered list 12, ..., 15, 23, ..., 25, 34, 35, 45 (using the notation ij to denote the interaction A_iA_j). However, these cases occurred very rarely in our simulation, and hence this step was rarely employed. In real life, investigators can come up with rules to proceed based on additional analysis (see [88], for an example).

The abstract discussion of Algorithm 1 and refining phase possibilities are made more concrete below with the help of three simulated examples:

Example 1

Suppose at the screening phase, the significant effects along with their signs are:

$$A_1(+), A_2(+), A_3(-)$$

Before running any follow-up experiment, Algorithm 1 will operate on this as follows:

- Best combination is initialized as $[0, 0, 0, 0, 0]$.

- By Step 1, A_1, A_2, A_3, A_4 and A_5 are set to 1, 1, 0, 0, and 0 respectively. So Best combination becomes $[1, 1, 0, 0, 0]$.

Since only the main effect of A_1 (but none of its interactions) is significant at the screening phase, a 2-group follow-up experiment is conducted at the refining phase, where the 2 groups correspond to the levels 1 and 2 of A_1. In both groups, A_2, A_3, A_4, and A_5 are fixed at levels 1, 0, 0, and 0 respectively – as determined by Algorithm 1 above.

Example 2

Suppose at the screening phase, the significant effects along with their signs are:

$$A_1(+), A_2(+), A_4(+), A_1A_2(+), A_4A_5(-)$$

Before running any follow-up experiment, Algorithm 1 will operate on this as follows:

- Best combination is initialized as $[0, 0, 0, 0, 0]$.
- By Step 1, A_3 is set to 0. Best combination remains same as before.
- By Step 2 (a – c), *sign vector* for the interaction A_1A_2 is $(1, 1, 1)$, and the *sign vector* for the interaction A_4A_5 is $(1, -1, -1)$.
- By Step 3, best combination becomes $[1, 1, 0, 1, 0]$.

Since one interaction involving A_1 is significant at the screening phase, a 6-group follow-up experiment is conducted at the refining phase, where the 6 groups correspond to the combinations (0, 0), (0, 1), (1, 0), (1, 1), (2, 0), and (2, 1) of the components A_1 and A_2. In all 6 groups, A_3, A_4, and A_5 are fixed at levels 0, 1, and 0 respectively – as determined by Algorithm 1 above.

Example 3

Suppose at the screening phase, the significant effects along with their signs are:

$$A_1(+), A_2(+), A_5(+), A_2A_3(-), A_3A_5(+)$$

Before running any follow-up experiment, Algorithm 1 will operate on this as follows:

- Best combination is initialized as $[0, 0, 0, 0, 0]$.
- By Step 1, A_1 is set at 1 and A_4 is set at 0. Best combination becomes $[1, 0, 0, 0, 0]$.
- By Step 2 (a – c), *sign vector* for the interaction A_2A_3 is $(1, -1, -1)$, and the *sign vector* for the interaction A_3A_5 is $(1, 1, 1)$.
- Step 3 applied to the interaction A_2A_3 sets A_2 at 1 and A_3 at 0. Best combination becomes $[1, 1, 0, 0, 0]$.
- Step 3 applied to the interaction A_3A_5 sets A_3 at 1 and A_5 at 1. Thus the best combination becomes $[1, 1, 1, 0, 1]$.

This is an example where the order in which interactions are considered in Algorithm 1 affects the results. Since in our simulations, we considered natural ordering as described in Algorithm 1, A_3A_5 is considered after A_2A_3. We could have ended up with a different best combination (e.g., $[1, 1, 0, 0, 1]$) had we considered A_3A_5 before A_2A_3.

Since only the main effect of A_1 (but none of its interactions) is significant at the screening phase, a 2-group follow-up experiment is conducted at the refining phase, where the 2 groups correspond to the levels 1 and 2 of A_1. In both groups, A_2, A_3, A_4, and A_5 are fixed at levels 1, 1, 0, and 1 respectively – as determined by Algorithm

1 above.

The matlab code for the entire simulation can be found at

`http://www.stat.lsa.umich.edu/~bibhas/MOSTcode.html`

4.7 Appendix C: Summary Results across Different Simulation Conditions

We conducted a series of simulations in order to investigate whether the results reported in the main text held across variation along two dimensions: (1) sample size (N), and (2) number of largest main effects retained in the screening phase as a decision rule (k). We investigated three different sample sizes: $N = 600$ (400+200), 1200 (800+400), and 2500 (1600+900); crossed with three different decision rules: $k = 2$, 3, and 4 (i.e., nine simulation settings in total). The following three tables correspond to the three tables in the main text, summarizing across all the nine settings. In general the classical approach tended to produce a larger $E(Y)$ than the MOST approach in small effect size, small sample size conditions; and the MOST approach tended to produce a larger $E(Y)$ than the classical approach in medium or better effect size conditions, even with small sample sizes ($N = 600$). The MOST approach tended to produce larger $E(Y)$ than the classical approach for larger k. In the conditions in which the four largest main effects were retained, the MOST approach consistently produced the larger $E(Y)$, even in the small effect size, small sample size ($N = 600$) condition. Details can be seen in the following tables.

Table 4.4: Whether the Classical (C) or the MOST (M) approach produced the largest value of $E(Y)$ under a variety of simulation conditions (This is a summary across 9 simulation settings of the entries that correspond to the Table 1 in the main text).

		k		
Sample Size	Effect Size	$k = 2$	$k = 3$	$k = 4$
600	Small	C	C	M
	Medium	C	M	M
	Large	M	M	M
1200	Small	C	C	M
	Medium	M	M	M
	Large	M	M	M
2500	Small	C	M	M
	Medium	M	M	M
	Large	M	M	M

Table 4.5: Whether the Classical (C) or the MOST (M) approach produced a higher $E(Y)$ value in more data sets than its counterpart under a variety of simulation conditions (This is a summary across 9 simulation settings of the entries that correspond to the Table 2 in the main text).

		k		
Sample Size	Effect Size	$k = 2$	$k = 3$	$k = 4$
600	Small	C	C	M
	Medium	C	M	M
	Large	M	M	M
1200	Small	C	C	M
	Medium	M	M	M
	Large	M	M	M
2500	Small	C	M	M
	Medium	M	M	M
	Large	M	M	M

Table 4.6: Whether the Classical (C) or the MOST (M) approach showed more accuracy in component selection under a variety of simulation conditions (This is a summary across 9 simulation settings of the entries that correspond to the Table 3 in the main text).

Dimension	Sample Size	Effect Size	k		
			$k=2$	$k=3$	$k=4$
Identifying the correct combination of components and levels more frequently	600	Small	C	M	M
		Medium	M	M	M
		Large	M	M	M
	1200	Small	M	M	M
		Medium	M	M	M
		Large	M	M	M
	2500	Small	M	M	M
		Medium	M	M	M
		Large	M	M	M
Identifying all the active components more frequently	600	Small	C	C	C
		Medium	C	C	M
		Large	C	M	M
	1200	Small	C	C	C
		Medium	C	C	M
		Large	M	M	M
	2500	Small	C	C	C
		Medium	M	M	M
		Large	M	M	M
Identifying all the inactive components more frequently	600	Small	M	M	M
		Medium	M	M	M
		Large	M	M	M
	1200	Small	M	M	M
		Medium	M	M	M
		Large	M	M	M
	2500	Small	M	M	M
		Medium	M	M	M
		Large	M	M	M

CHAPTER V

Inference for Non-regular Parameters in Optimal Dynamic Treatment Regimes

This chapter provides a detailed treatment of the problem of non-regularity in the context of optimal dynamic treatment regimes. In the estimation of the optimal dynamic treatment regime from longitudinal data, the treatment effect parameters at any stage prior to the last can be non-regular under certain distributions of the data. This results in biased estimates and invalid confidence intervals for the treatment effect parameters. In this chapter, we discuss both the problem of non-regularity, and available estimation methods. We provide an extensive simulation study to compare the estimators in terms of their ability to lead to valid confidence intervals under a variety of non-regular scenarios. Analysis of the smoking cessation trial data is provided as an illustration.

5.1 Introduction

Many diseases such as mental illnesses, HIV infection, and substance abuse are clinically treated in multiple stages, adapting the treatment type and dosage to the ongoing measures of an individual patient's response, adherence, burden, side effects, and preference. Dynamic treatment regimes represent one way to operationalize this

sequential decision making. A dynamic treatment regime (DTR) is a sequence of decision rules, one per stage. Each decision rule takes a patient's treatment and covariate history as input, and outputs a recommended treatment. The main motivations for considering sequences of treatments are high variability across patients in response to any one type of treatment, likely relapse, presence or emergence of co-morbidities, time-varying side effect severity, and reduction of costs and burden when intensive treatment is unnecessary [24].

A DTR is said to be optimal if it optimizes the mean outcome at the end of the final stage of treatment. Data for estimating the optimal DTR can come from either an observational longitudinal study or a sequential multiple assignment randomized trial (SMART) [49, 50, 32, 57]. In these designs, each patient is followed through stages of treatment and at each stage the patient is randomized to one of the possible treatment options. Experimental designs similar to SMART have been implemented in the treatments of schizophrenia [81], depression [78], and cancer [93, 99].

Estimating the optimal DTR is a problem of sequential, multi-stage decision making. Murphy [56] developed a semiparametric method for estimating the optimal DTR, an efficient version of which was provided by Robins [71]. A good discussion of the relationship between these two methods can be found in Moodie et al. [55]. Other methods for estimating optimal DTRs in the literature include likelihood-based methods, both frequentist and Bayesian, developed by Thall et al. [93, 94, 95], and the semiparametric methods of Lunceford et al. [51], and Wahed and Tsiatis [99, 100].

Robins [71] considered the problem of inference for the parameters of the optimal DTR. As discussed by Robins, the treatment effect parameters at any stage prior to the last can be *non-regular* under certain longitudinal distributions of the data

which he called *exceptional laws*. By non-regularity, we mean that the asymptotic distribution of the estimator of the treatment effect parameter does not converge uniformly over the parameter space (see below for further details). This phenomenon of non-regularity causes bias in estimation, and leads to poor frequentist properties of the confidence intervals. Recently Moodie and Richardson [54] provided a method called *Zeroing Instead of Plugging In* (ZIPI) for correcting the bias in the estimation of the optimal DTRs resulting under exceptional laws.

The main goals of this chapter are to illustrate the problem of non-regularity, and to compare available estimation methods that attempt to address this problem. In section 5.2, we discuss the problem of non-regularity in detail. Section 5.3 provides a description of different methods that address the problem. We provide an extensive simulation study in section 5.4 to compare the estimators in terms of their ability to lead to valid confidence intervals using bootstrap. This is followed by an analysis of a data set from a longitudinal smoking cessation trial in section 5.5; the purpose is to demonstrate the applicability of the estimation methods in a real-life non-regular scenario. Finally an overall discussion is provided in section 5.6. Throughout this chapter, we assume that the data come from SMART designs. The main reason for this is to separate the issue of non-regularity from causal inference issues. However the problem of non-regularity also arises when observational data are used [71, 54]; and the estimators proposed in section 5.3 should be applicable to observational data as well. Technical details are presented in the appendices at the end of the chapter.

5.2 Estimation and Inference via Q-learning

5.2.1 Notation and Data Structure

For simplicity, we focus on studies with two stages. Longitudinal data on a single patient are given by the trajectory $(O_1, A_1, O_2, A_2, O_3)$, where O_j $(j = 1, 2)$ denotes the covariates measured prior to treatment at the beginning of the j-th stage, O_3 is the observation at the end of stage 2, and A_j $(j = 1, 2)$ is the treatment assigned at the j-th stage subsequent to observing O_j. The data set consists of a random sample of n patients. Define the history at each stage as: $H_1 = O_1,\ H_2 = (O_1, A_1, O_2)$. We consider a SMART design in which there are two possible treatments at each stage, $A_j \in \{-1, 1\}$; here we assume $P[A_j = -1|H_j] = P[A_j = 1|H_j] = \frac{1}{2}$. The study can have either a single primary outcome Y observed at the end of stage 2, or two outcomes Y_1, Y_2 observed at the two stages. Note that the case of a single outcome Y observed at the end can be viewed as a case with $Y_1 \equiv 0$ and $Y_2 = Y$. We assume $Y_1 = f_1(O_1, A_1, O_2)$ and $Y_2 = f_2(O_1, A_1, O_2, A_2, O_3)$, with known functions f_1, f_2. A two-stage DTR consists of two decision rules, say (d_1, d_2), with $d_j(H_j) \in \mathcal{A}_j$, where $\mathcal{A}_j$ is the set of possible treatments at the j-th stage.

One simple method to construct (d_1, d_2) is Q-learning [101, 90, 58]. Q-learning, like Robins' *g-estimation of optimal structural nested mean models* (hereafter simply referred to as *Robins' method*), suffers from non-regularity – the common reason being an underlying non-smooth maximization operation. Here we will illustrate the problem due to non-regularity using Q-learning, since it can be viewed as a generalization of the least squares regression to multistage decision problems, and hence simpler to explain than Robins' semiparametric efficient method. In Lemma V.1 below, we provide conditions under which Q-learning is equivalent to an inefficient version of Robins' method.

5.2.2 Q-learning with Linear Models

First let us define the Q-functions [90, 58] for the two stages as follows:

$$\begin{aligned} Q_2(H_2, A_2) &= E\big[Y_2|H_2, A_2\big], \\ Q_1(H_1, A_1) &= E\big[Y_1 + \max_{a_2} Q_2(H_2, a_2)|H_1, A_1\big]. \end{aligned}$$

If the two Q-functions were known, the optimal DTR (d_1, d_2), using backwards induction (as in dynamic programming) argument, would be

$$d_j(h_j) = \arg\max_{a_j} Q_j(h_j, a_j), \quad j = 1, 2. \tag{5.1}$$

In practice, the true Q-functions are not known and hence must be estimated from the data. Consider a linear model for the Q-functions. Let the stage-j $(j = 1, 2)$ Q-function be modeled as

$$Q_j(H_j, A_j; \beta_j, \psi_j) = \beta_j^T H_{j0} + (\psi_j^T H_{j1}) A_j, \tag{5.2}$$

where H_{j0} and H_{j1} are two (possibly different) summaries of the history H_j, with H_{j0} denoting the "main effect of history" and H_{j1} denoting the part of history that interacts with treatment (H_{j0} and H_{j1} include the intercept term). The Q-learning algorithm is:

1. Stage-2 regression: $(\hat{\beta}_2, \hat{\psi}_2) = \arg\min_{\beta_2, \psi_2} \frac{1}{n} \sum_{i=1}^{n} \Big(Y_{2i} - Q_2(H_{2i}, A_{2i}; \beta_2, \psi_2)\Big)^2$.
2. Stage-2 optimal rule: $\hat{d}_2(h_2) = \arg\max_{a_2} Q_2(h_2, a_2; \hat{\beta}_2, \hat{\psi}_2)$.
3. Stage-1 pseudo-outcome: $\hat{Y}_{1i} = Y_{1i} + \max_{a_2} Q_2(H_{2i}, a_2; \hat{\beta}_2, \hat{\psi}_2)$, $i = 1, \ldots, n$.
4. Stage-1 regression: $(\hat{\beta}_1, \hat{\psi}_1) = \arg\min_{\beta_1, \psi_1} \frac{1}{n} \sum_{i=1}^{n} \Big(\hat{Y}_{1i} - Q_1(H_{1i}, A_{1i}; \beta_1, \psi_1)\Big)^2$.
5. Stage-1 optimal rule: $\hat{d}_1(h_1) = \arg\max_{a_1} Q_1(h_1, a_1; \hat{\beta}_1, \hat{\psi}_1)$.

The estimated optimal DTR using Q-learning is given by $(\hat{d}_1, \hat{d}_2)$.

The following lemma gives a set of sufficient conditions under which Q-learning is equivalent to an inefficient version of Robins' method.

Lemma V.1. *Consider linear models for the Q-functions as in (5.2). Assume that:*

(i) the parameters in Q_1 and Q_2 are distinct;

(ii) A_j has zero conditional mean given the history H_j, $j = 1, 2$; and

(iii) the covariates used in the model for Q_1 are nested within the covariates used in the model for Q_2, i.e., $(H_{10}^T, H_{11}^T A_1) \subset H_{20}^T$.

Then Q-learning is algebraically equivalent to an inefficient version of Robins' method.

The proof is given in Appendix A.

5.2.3 The Inference Problem

With (5.2) as the model for Q-functions, the optimal DTR is given by

$$d_j(H_j) = \arg\max_{a_j} (\psi_j^T H_{j1}) a_j = sign(\psi_j^T H_{j1}), \quad j = 1, 2, \tag{5.3}$$

where $sign(x) = 1$ if $x > 0$, and -1 otherwise. Note that the term $\beta_j^T H_{j0}$ on the right side of (5.2) does not feature in the optimal DTR. Thus for estimating optimal DTRs, the ψ_j's are the parameters of interest, while β_j's are nuisance parameters. We want to perform inference (e.g., construct confidence intervals) on ψ_j's.

Conducting inference on ψ_j's is important for the following reasons. First, if the confidence intervals (or hypothesis tests) for ψ_j reveal that there is evidence that some components of the parameter vector ψ_j are not clinically different from zero,

then the investigator may choose not to collect the corresponding components of the history vector H_{j1} while making decisions using the optimal DTR. This reduces the cost of data collection in a future implementation of the optimal DTR. Thus in the present context, confidence intervals (or hypothesis tests) can be viewed as a tool for doing variable selection. Second, it is important to know when there is insufficient support in the data to recommend one treatment over another, since in such cases treatment can be chosen according to other considerations like cost, familiarity, burden, preference etc. Third, as discussed by Robins [71], confidence intervals for ψ_j can lead to confidence intervals for d_j. In the following, we discuss the problem of non-regularity in inference.

5.2.4 Non-regularity in Inference

Note that the stage-1 pseudo-outcome (in the Q-learning algorithm) is

$$\hat{Y}_{1i} = Y_{1i} + \max_{a_2} Q_2(H_{2i}, a_2; \hat{\beta}_2, \hat{\psi}_2) = Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}|, \tag{5.4}$$

which is a non-smooth (e.g., non-differentiable at $\hat{\psi}_2^T H_{21,i} = 0$) function of $\hat{\psi}_2$, because of the maximization operation. Since $\hat{\psi}_1$ is a function of $\hat{Y}_{1i}$, $i = 1, \ldots, n$, it is in turn a non-smooth function of $\hat{\psi}_2$. As a consequence, the asymptotic distribution of $\sqrt{n}(\hat{\psi}_1 - \psi_1)$ does not converge uniformly [71] over the parameter space of $\psi = (\psi_1, \psi_2)$. More specifically, the asymptotic distribution of $\sqrt{n}(\hat{\psi}_1 - \psi_1)$ is normal if ψ_2 is such that $P[H_2 : \psi_2^T H_{21} = 0] = 0$, but is non-normal if $P[H_2 : \psi_2^T H_{21} = 0] > 0$. This change in the asymptotic distribution happens abruptly. The (vector) parameter ψ_1 is called a *non-regular* parameter and the estimator $\hat{\psi}_1$ is called a *non-regular* estimator; see [5] for the precise definition of non-regularity. Because of this non-regularity, given the noise level present in small samples, the estimator $\hat{\psi}_1$ oscillates

between the two asymptotic distributions across samples. As a result, usual Wald type confidence intervals perform poorly [71, 54].

The issue of non-regularity can be better understood with a toy example discussed by Robins [71] (here is a slightly modified version). Consider the problem of estimating $|\mu|$ based on n i.i.d. observations $X_1, \ldots, X_n$ from $N(\mu, 1)$. Note that $|\bar{X}_n|$ is the maximum likelihood estimator of $|\mu|$, where $\bar{X}_n$ is the sample average. It can be shown that the asymptotic distribution of $\sqrt{n}(|\bar{X}_n| - |\mu|)$ for $\mu = 0$ is different from that for $\mu \neq 0$. Thus $|\bar{X}_n|$ is a non-regular estimator of $|\mu|$. Also, for $\mu = 0$, $\lim_{n\to\infty} E[\sqrt{n}(|\bar{X}_n| - |\mu|)] = \sqrt{\frac{2}{\pi}}$. Robins referred to this quantity as the *asymptotic bias* of the estimator $|\bar{X}_n|$. This asymptotic bias is one symptom of the underlying non-regularity, as discussed by Moodie and Richardson [54].

In many situations where the asymptotic distribution of an estimator is unavailable, bootstrap is used as an alternative approach to conduct inference. But the success of bootstrap also hinges on the underlying smoothness of the estimator. When an estimator is non-smooth, the ordinary (n out of n) bootstrap procedure produces an inconsistent bootstrap estimator [84]. Inconsistency of bootstrap in the above simple normal theory example has been discussed by Andrews [2]. As shown by Shao [84], an alternative resampling procedure called "m out of n bootstrap" is consistent in such non-smooth scenarios. One concern regarding the use of this procedure is the slower rate of convergence than $\sqrt{n}$ even in a regular setting (e.g., when $P[H_2 : \psi_2^T H_{21} = 0] = 0$). Moreover, a data-adaptive choice of the tuning parameter m in the present context of DTRs is not obvious; see however [6] and [43] for data-adaptive choice of m in other contexts.

The above concerns regarding non-regularity led us to investigate possible regularizations of the estimation procedure, and then use bootstrap for inference. In

the simulation study to follow, we will investigate the behavior of different types of bootstrap confidence intervals for the parameters ψ_j of the optimal DTR in both regular and non-regular settings.

5.3 Different Regularized Estimators

In this section, we will present two competing estimators to address the non-regularity problem described above. Limited theoretical results are available at this point, and consequently it is not clear which estimator is better. In this chapter, we will study their relative merits and demerits in simulations.

From the discussion on non-regularity above, it is clear that $\hat{\psi}_1$ is a non-regular estimator because the stage-1 pseudo-outcome $\hat{Y}_1$ is a non-smooth function (e.g., absolute value) of $\hat{\psi}_2$. The estimators presented in this section "regularize" the non-regular estimator (sometimes called the "hard-max" estimator because of the maximum operation used in the definition) by shrinking or thresholding the effect of the term involving the maximum, e.g., $|\hat{\psi}_2^T H_{21}|$, towards zero.

5.3.1 Hard-threshold Estimator

Recall that the pseudo-outcome $\hat{Y}_1 = Y_1 + \hat{\beta}_2^T H_{20} + |\hat{\psi}_2^T H_{21}|$ is non-differentiable in $\hat{\psi}_2$ only when $\hat{\psi}_2^T H_{21} = 0$, and so the corresponding estimator $\hat{\psi}_1$ is problematic only when the true $\psi_2^T H_{21}$ is close to zero. The general form of the hard-threshold pseudo-outcome is

$$\hat{Y}_{1i}^{HT} = Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}| \cdot \mathbf{1}\{|\hat{\psi}_2^T H_{21,i}| > \lambda_i\}, \tag{5.5}$$

where λ_i (> 0) is the threshold for the i-th subject in the sample (possibly depending on the variability of the linear combination $\hat{\psi}_2^T H_{21,i}$ for that subject). One way to

operationalize this is to perform a preliminary test (for each subject in the sample) of the hypothesis $H_{0i} : \psi_2^T H_{21,i} = 0$ ($H_{21,i}$ is considered fixed in this test), set $\hat{Y}_{1i}^{HT} = \hat{Y}_{1i}$ if H_{0i} is rejected, and replace $|\hat{\psi}_2^T H_{21,i}|$ with the "better guess" 0 in case H_{0i} is accepted. Thus the hard-threshold pseudo-outcome can be written as

$$\hat{Y}_{1i}^{HT} = Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}| \cdot \mathbf{1}\left\{ \frac{\sqrt{n}|\hat{\psi}_2^T H_{21,i}|}{\sqrt{H_{21,i}^T \hat{\Sigma}_2 H_{21,i}}} > z_{\alpha/2} \right\}, \tag{5.6}$$

where $\hat{\Sigma}_2$ is the estimated covariance matrix of $\hat{\psi}_2$. The corresponding estimator of ψ_1, denoted by $\hat{\psi}_1^{HT}$, will be referred to as the hard-threshold estimator. The hard-threshold estimator is common in many areas like variable selection in linear regression and wavelet shrinkage [34]. Moodie and Richardson [54] proposed this estimator for bias correction in the context of Robins' method, and called it *Zeroing Instead of Plugging In* (ZIPI) estimator.

Note that $\hat{Y}_1^{HT}$ is still a non-smooth function of $\hat{\psi}_2$ and hence $\hat{\psi}_1^{HT}$ is a non-regular estimator of ψ_1. However, the problematic term $|\hat{\psi}_2^T H_{21}|$ is shrunk (thresholded) towards zero, and hence one might expect that the degree of non-regularity is somewhat reduced. Moodie and Richardson[54] showed that this estimator reduces the bias occurring in Robins' method (efficient version of Q-learning). In the simulation study to follow, we will explore if this estimator can be used to construct valid confidence intervals for ψ_1. An important issue regarding the use of this estimator is the choice of significance level α of the preliminary test, which is an unknown tuning parameter. As discussed by Moodie and Richardson[54], this is a difficult problem even in better-understood settings where preliminary test based estimators are used; and no widely applicable data-driven method for choosing α in this setting is currently available.

5.3.2 Soft-threshold or Shrinkage Estimator

The general form of the soft-threshold pseudo-outcome considered here is

$$\hat{Y}_{1i}^{ST} = Y_{1i} + \hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}| \cdot \left(1 - \frac{\lambda_i}{|\hat{\psi}_2^T H_{21,i}|^2}\right)^+, \tag{5.7}$$

where $x^+ = x\mathbf{1}\{x > 0\}$ stands for the positive part of a function, and λ_i (> 0) is a tuning parameter associated with the i-th subject in the sample (again possibly depending on the variability of the linear combination $\hat{\psi}_2^T H_{21,i}$ for that subject). In the contexts of regression shrinkage [12] and wavelet shrinkage [40], the third term in (5.7) is generally known as the *nonnegative garrote* estimator. As discussed by Zou [109], the nonnegative garrote estimator is a special case of the *adaptive lasso* estimator. As in the case of hard-threshold estimator, a crucial issue here is to choose a data-driven tuning parameter λ_i. Below we provide a choice following a Bayesian approach.

Like the hard-threshold pseudo-outcome, $\hat{Y}_1^{ST}$ is also a non-smooth function of $\hat{\psi}_2$ and hence $\hat{\psi}_1^{ST}$ remains a non-regular estimator of ψ_1. However, the problematic term $|\hat{\psi}_2^T H_{21}|$ is shrunk (or thresholded) towards zero, and hence one might expect that the degree of non-regularity is somewhat reduced. In the simulation study to follow, we will investigate how much improvement this estimator offers over the "hard-max" estimator, when it comes to constructing confidence intervals. Figure 1 presents the hard-max, the hard-threshold, and the soft-threshold pseudo-outcomes.

Choice of Tuning Parameter

A hierarchical Bayesian formulation of the problem, inspired by the work of Figueiredo and Nowak [35] in the area of wavelet-based image processing, can be

used in the context of the soft-threshold estimator to choose λ_i's in a data-driven way. It turns out that the estimator (5.7) with $\lambda_i = 3H_{21,i}^T \hat{\Sigma}_2 H_{21,i}/n,\ i = 1, \ldots, n$, where $\hat{\Sigma}_2/n$ is the estimated covariance matrix of $\hat{\psi}_2$, is an approximate empirical Bayes estimator. The following lemma will be used to derive the choice of λ_i.

Lemma V.2. *Let X be a random variable such that $X|\mu \sim N(\mu, \sigma^2)$ with known variance σ^2. Let the prior distribution on μ be given by $\mu|\phi^2 \sim N(0, \phi^2)$, with Jeffrey's noninformative hyper-prior on ϕ^2, e.g., $p(\phi^2) \propto 1/\phi^2$. Then an empirical Bayes estimator of $|\mu|$ is given by*

$$\begin{aligned} \widehat{|\mu|}^{EB} &= X\left(1-\frac{3\sigma^2}{X^2}\right)^+\left(2\Phi\Big(\frac{X}{\sigma}\sqrt{\Big(1-\frac{3\sigma^2}{X^2}\Big)^+}\Big)-1\right) \\ &+ \sqrt{\frac{2}{\pi}}\,\sigma\sqrt{\Big(1-\frac{3\sigma^2}{X^2}\Big)^+}\exp\left\{-\frac{X^2}{2\sigma^2}\left(1-\frac{3\sigma^2}{X^2}\right)^+\right\}, \end{aligned} \tag{5.8}$$

where $\Phi(\cdot)$ is the standard normal distribution function.

The proof is given in Appendix B.

Clearly, $\widehat{|\mu|}^{EB}$ is a thresholding rule, since $\widehat{|\mu|}^{EB} = 0$ for $|X| < \sqrt{3}\sigma$. Moreover,

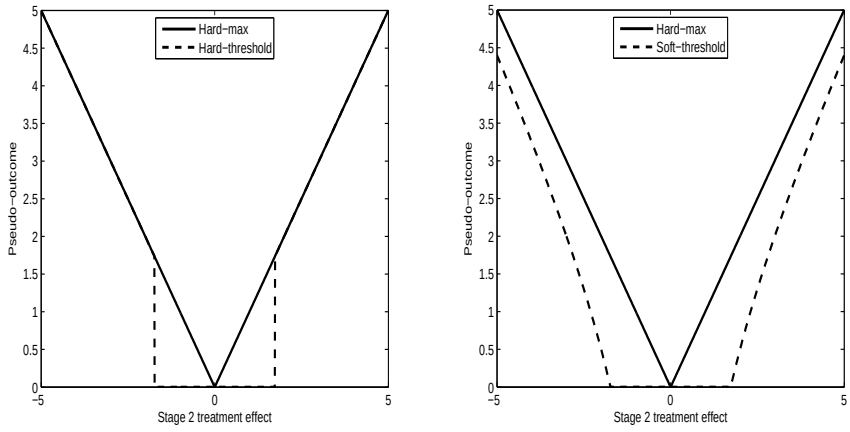

Figure 5.1: Hard-threshold and Soft-threshold pseudo-outcomes compared with the Hard-max pseudo-outcome.

when $|\frac{X}{\sigma}|$ is large, the second term of (5.8) goes to zero exponentially fast, and

$$\left(2\Phi\Big(\frac{X}{\sigma}\sqrt{\Big(1-\frac{3\sigma^2}{X^2}\Big)^+}\Big)-1\right)\approx(2\,I_{\{X>0\}}-1)=sign(X).$$

Consequently, the empirical Bayes estimator is approximated by

$$\widehat{|\mu|}^{EB}\approx X\left(1-\frac{3\sigma^2}{X^2}\right)^+ sign(X)=|X|\left(1-\frac{3\sigma^2}{X^2}\right)^+. \tag{5.9}$$

Now for $i=1,\ldots,n$ separately, put $X=\hat{\psi}_2^T H_{21,i}$, and $\mu=\psi_2^T H_{21,i}$ (for fixed $H_{21,i}$); and plug in $\hat{\sigma}^2=H_{21,i}^T\hat{\Sigma}_2 H_{21,i}/n$ for σ^2. This leads to a choice of λ_i in the soft-threshold pseudo-outcome (5.7):

$$\begin{aligned}\hat{Y}_{1i}^{ST} &= Y_{1i}+\hat{\beta}_2^T H_{20,i}+|\hat{\psi}_2^T H_{21,i}|\cdot\left(1-\frac{3H_{21,i}^T\hat{\Sigma}_2 H_{21,i}}{n|\hat{\psi}_2^T H_{21,i}|^2}\right)^+,\\ &= Y_{1i}+\hat{\beta}_2^T H_{20,i}+|\hat{\psi}_2^T H_{21,i}|\cdot\left(1-\frac{3H_{21,i}^T\hat{\Sigma}_2 H_{21,i}}{n|\hat{\psi}_2^T H_{21,i}|^2}\right)\cdot\mathbf{1}\left\{\frac{\sqrt{n}|\hat{\psi}_2^T H_{21,i}|}{\sqrt{H_{21,i}^T\hat{\Sigma}_2 H_{21,i}}}>\sqrt{3}\right\},\\ &\qquad\qquad i=1,\ldots,n.\end{aligned} \tag{5.10}$$

The presence of the indicator function in (5.10) indicates that $\hat{Y}_{1i}^{ST}$ is a thresholding rule for small values of $|\hat{\psi}_2^T H_{21,i}|$, while the term just preceding the indicator function makes $\hat{Y}_{1i}^{ST}$ a shrinkage rule for moderate to large values of $|\hat{\psi}_2^T H_{21,i}|$ (for which the indicator function takes the value one). Thus the current Bayesian formulation gives us a data-driven choice of the tuning parameters.

5.4 Simulation Study

In this section, we consider a simulation study to compare the performances of the hard-max, the hard-threshold, and the soft-threshold estimators under different non-regular scenarios. In this study, we vary the parameters of the generative model, the degree of non-regularity, and the type of bootstrap confidence interval.

Generative Model

Recall that the data consist of n patient trajectories. Each such trajectory is of the form $(O_1, A_1, O_2, A_2, O_3)$. Without loss of generality, we assume $Y_1 \equiv 0$ and $Y_2 \equiv Y = O_3$. Let $\mu_Y = E[Y|O_1, A_1, O_2, A_2]$, and ϵ be the associated error term. Then $Y = \mu_Y + \epsilon$, where

$$\mu_Y = \gamma_1 + \gamma_2 O_1 + \gamma_3 A_1 + \gamma_4 O_1 A_1 + \gamma_5 A_2 + \gamma_6 O_2 A_2 + \gamma_7 A_1 A_2,$$

and $\epsilon \sim N(0, 1)$. Next, we consider binary treatments randomized with probability 1/2, e.g., $P[A_j = 1] = P[A_j = -1] = 1/2,\ j = 1, 2$. Also, the binary covariates O_j's are generated as

$$\begin{aligned} P[O_1 = 1] &= P[O_1 = -1] = 1/2, \\ P[O_2 = 1|O_1, A_1] &= 1 - P[O_2 = -1|O_1, A_1] = expit(\delta_1 O_1 + \delta_2 A_1), \end{aligned}$$

where $expit(x) = \exp(x)/(1 + \exp(x))$. Note that $\gamma_1, \ldots, \gamma_7$ and δ_1, δ_2 are the parameters that specify the generative model. These parameters will be varied in the examples to follow.

Analysis Model

$$\begin{aligned} Q_2(H_2, A_2) &= \beta_{20} + \beta_{21} O_1 + \beta_{22} A_1 + \beta_{23} O_1 A_1 + \Big(\psi_{20} + \psi_{21} O_2 + \psi_{22} A_1\Big) A_2, \\ Q_1(H_1, A_1) &= \beta_{10} + \beta_{11} O_1 + \Big(\psi_{10} + \psi_{11} O_1\Big) A_1. \end{aligned}$$

Two dimensions of non-regularity: p and ϕ

Non-regularity in stage 1 parameters arises when the optimal stage 2 treatment is non-unique for at least some subjects in the population. With reference to the present generative model, a setting is non-regular if the linear combination $\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1 =$

Table 5.1: Distribution of the linear combination $(\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1)$

(O_2, A_1) cell	cell probability (averaged over O_1)	value of the linear combination
$(1,1)$	$q_1 \equiv \frac{1}{4}\Big(expit(\delta_1+\delta_2) + expit(-\delta_1+\delta_2)\Big)$	$f_1 \equiv \gamma_5+\gamma_6+\gamma_7$
$(1,-1)$	$q_2 \equiv \frac{1}{4}\Big(expit(\delta_1-\delta_2) + expit(-\delta_1-\delta_2)\Big)$	$f_2 \equiv \gamma_5+\gamma_6-\gamma_7$
$(-1,1)$	$q_3 \equiv \frac{1}{4}\Big(expit(\delta_1-\delta_2) + expit(-\delta_1-\delta_2)\Big)$	$f_3 \equiv \gamma_5-\gamma_6+\gamma_7$
$(-1,-1)$	$q_4 \equiv \frac{1}{4}\Big(expit(\delta_1+\delta_2) + expit(-\delta_1+\delta_2)\Big)$	$f_4 \equiv \gamma_5-\gamma_6-\gamma_7$

0 with positive probability. Also one might expect some non-regular behavior as $\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1$ falls in a small neighborhood of zero (even though not exactly zero). In the following, we consider specific examples varying the "degree of non-regularity", e.g., $p = P[\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1 = 0]$ and the "standardized effect size" defined as $\phi = \Big|E[\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1]/\sqrt{Var[\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1]}\Big|$. The quantities p and ϕ, which depend on the distribution of the above linear combination, represent two dimensions of the non-regularity phenomenon. Note that the linear combination $(\gamma_5 + \gamma_6 O_2 + \gamma_7 A_1)$ can take only four possible values corresponding to the four possible (O_2, A_1) cells. The cell probabilities can be easily calculated; the formulae are provided in Table 5.1.

It follows that $E[\gamma_5+\gamma_6 O_2+\gamma_7 A_1] = q_1 f_1 + q_2 f_2 + q_3 f_3 + q_4 f_4$, and $E[(\gamma_5+\gamma_6 O_2+\gamma_7 A_1)^2] = q_1 f_1^2 + q_2 f_2^2 + q_3 f_3^2 + q_4 f_4^2$, where $q_1, \ldots, q_4$ are the cell probabilities given in Table 1. From these two, one can calculate $Var[\gamma_5+\gamma_6 O_2+\gamma_7 A_1]$, and subsequently the effect size ϕ.

We want to conduct inference on ψ_{10} and ψ_{11}, the analysis model parameters associated with stage 1 treatment A_1. They can be expressed in terms of γ's and δ's, the parameters of the generative model, as follows. It turns out that

$$\begin{aligned} \psi_{10} &= \gamma_3 + q_1|f_1| - q_2|f_2| + q_3|f_3| - q_4|f_4|, \\ \text{and } \psi_{11} &= \gamma_4 + q_1'|f_1| - q_2'|f_2| - q_3'|f_3| + q_4'|f_4|, \end{aligned}$$

where $q'_1 = q'_3 = \frac{1}{4}(expit(\delta_1 + \delta_2) - expit(-\delta_1 + \delta_2))$, and $q'_2 = q'_4 = \frac{1}{4}(expit(\delta_1 - \delta_2) - expit(-\delta_1 - \delta_2))$. In the following, we consider specific examples for varying p and ϕ. In Examples $1-4$ below, we use $\delta_1 = \delta_2 = 0.5$. For this choice, we get the following values of the cell probabilities: $q_1 = q_4 = 0.3078$ and $q_2 = q_3 = 0.1922$. This choice of the δ's also makes $q'_1 = q'_2 = q'_3 = q'_4 = .0578$.

Example 1 ($p = 1$, ϕ undefined): Consider a setting where there is no treatment effect for any subject (any history) in either stage. This is achieved by setting $\gamma_1 = \ldots = \gamma_7 = 0$, and $\delta_1 = \delta_2 = 0.5$. Then $f_1 = f_2 = f_3 = f_4 = 0$, and hence $\psi_{10} = \psi_{11} = 0$, $p = 1$, and ϕ is undefined $(0/0)$. This is a fully non-regular scenario.

Example 2 ($p = 0$, ϕ infinite): Consider a setting similar to Example 1, where there is a very weak stage 2 treatment effect for every subject (all possible history). This is achieved by setting $\gamma_5 = 0.01$ and $\gamma_j = 0, \forall j \neq 5$, and $\delta_1 = \delta_2 = 0.5$. Then $f_1 = f_2 = f_3 = f_4 = 0.01$; $\psi_{10} = \psi_{11} = 0$, $p = 0$, and ϕ is infinite $(0.01/0)$. This is a regular scenario, but close to non-regularity (it is hard to detect the very weak effect given the noise level in the data).

Example 3 ($p = \frac{1}{2}$, $\phi = 1$): Consider a setting where there is no stage 2 treatment effect for half the subjects in the population, but a reasonably large effect for the other half of subjects. This is achieved by setting $\gamma_1 = \gamma_2 = \gamma_4 = \gamma_6 = 0$, $\gamma_3 = -0.5$, $\gamma_5 = \gamma_7 = 0.5$, and $\delta_1 = \delta_2 = 0.5$. Then $f_1 = f_3 = 1$, $f_2 = f_4 = 0$, $\psi_{10} = \psi_{11} = 0$, $p = \frac{1}{2}$ and $\phi = 1$. This is a non-regular setting.

Example 4 ($p = 0$, $\phi = 1.0204$): Consider a setting where there is a very weak stage 2

treatment effect for half the subjects in the population, but a reasonably large effect for the other half of subjects. This is achieved by setting $\gamma_1 = \gamma_2 = \gamma_4 = \gamma_6 = 0$, $\gamma_3 = -0.5$, $\gamma_5 = 0.5$, $\gamma_7 = 0.49$, and $\delta_1 = \delta_2 = 0.5$. It follows that $f_1 = f_3 = 0.99$, $f_2 = f_4 = 0.01$, $\psi_{10} = -0.0100$, $\psi_{11} = 0$, $p = 0$, and $\phi = 1.0204$. This regular example is close to the non-regular Example 3.

Example 5 ($p = \frac{1}{4}$, $\phi = 1.4142$): Consider a setting where there is no stage 2 treatment effect for one-fourth of the subjects in the population, but others have a reasonably large effect. To achieve this, set $\gamma_1 = \gamma_2 = \gamma_4 = 0$, $\gamma_3 = -0.5$, $\gamma_5 = 1$, $\gamma_6 = \gamma_7 = 0.5$, $\delta_1 = 1$, and $\delta_2 = 0$. Then $f_1 = 2$, $f_2 = f_3 = 1$, $f_4 = 0$; the cell probabilities are equal, i.e., $q_1 = q_2 = q_3 = q_4 = \frac{1}{4}$; and $q_1' = q_2' = q_3' = q_4' = 0.1155$. Consequently, $\psi_{10} = \psi_{11} = 0$, $p = \frac{1}{4}$, and $\phi = 1.4142$. This is a non-regular setting.

Example 6 ($p = 0$, $\phi = 0.3451$): Consider a completely regular setting where there is a reasonably large stage 2 treatment effect for every subject in the population. This can be achieved by setting $\gamma_1 = \gamma_2 = \gamma_4 = 0$, $\gamma_3 = -0.5$, $\gamma_5 = 0.25$, $\gamma_6 = \gamma_7 = 0.5$, and $\delta_1 = \delta_2 = 0.1$. Then $f_1 = 1.25$, $f_2 = f_3 = 0.25$, and $f_4 = -0.75$; the cell probabilities are $q_1 = q_4 = 0.2625$, $q_2 = q_3 = 0.2375$; and $q_1' = q_2' = q_3' = q_4' = 0.0125$. It follows that $\psi_{10} = -0.3688$, $\psi_{11} = 0.0187$, $p = 0$ and $\phi = 0.3451$.

Note that in Example 5, the effect size ϕ is greater than Cohen's [21] benchmark large effect size (=0.8). Such a high effect size can be criticized as being unrealistic, based on the *principle of clinical equipoise* [38], which provides the ethical basis for medical research involving randomization. This principle says that there must be a honest, professional disagreement (high variability) among expert clinicians about

the preferred treatment (and thus the standardized effect size of treatment is likely small). Hence this example might be somewhat down-weighted for overall comparison of performance. Furthermore, Example 6 violates the *Hierarchical Ordering Principle* [107] in that the coefficient of the interaction term A_1A_2 (γ_7) is larger than the coefficient of the main effect A_2 (γ_5). So this example might be given lower weight as well.

Competing Estimators

In the simulation, we will consider four estimators: the hard-max estimator (original Q-learning), the soft-threshold estimator, and the hard-threshold estimator with two values of the tuning parameter α, e.g., 0.2, which was empirically found to be a good choice by Moodie and Richardson [54], and 0.08 which corresponds to the threshold used by the soft-threshold estimator proposed in this paper (from (5.10), the threshold used by the soft-threshold estimator is $\sqrt{3} = 1.7321$; equating this point to $z_{\alpha/2}$ and solving for α, we get $\alpha = 0.0833$).

Different Bootstrap CIs

We consider three types of bootstrap CIs, e.g., percentile, hybrid, and double (percentile) bootstrap CIs. Let $\hat{\theta}$ be an estimator of θ and $\hat{\theta}^*$ be its bootstrap version. Then the $100(1-\alpha)\%$ percentile bootstrap (PB) CI is given by $\left(\hat{\theta}^*_{(\frac{\alpha}{2})}, \hat{\theta}^*_{(1-\frac{\alpha}{2})}\right)$, and the $100(1-\alpha)\%$ hybrid bootstrap (HB) CI is given by $\left(2\hat{\theta} - \hat{\theta}^*_{(1-\frac{\alpha}{2})}, 2\hat{\theta} - \hat{\theta}^*_{(\frac{\alpha}{2})}\right)$, where $\hat{\theta}^*_\gamma$ is the 100γ-th percentile of the bootstrap distribution. The double bootstrap (DB) CI is calculated as follows:

1. Draw B_1 first-stage bootstrap samples from the original data. For each first-stage bootstrap sample, calculate the bootstrap version of the estimator $\hat{\theta}^{*b}$,

$b = 1, \ldots, B_1$.

2. Conditional on each first-stage bootstrap sample, draw B_2 second-stage (nested) bootstrap samples and calculate the double bootstrap versions of the estimator, e.g., $\hat{\theta}^{**bm},\ b = 1, \ldots, B_1,\ m = 1, \ldots, B_2$.

3. For $b = 1, \ldots, B_1$, calculate $u^{*b} = \frac{1}{B_2} \sum_{m=1}^{B_2} \mathbf{1}\{\hat{\theta}^{**bm} \leq \hat{\theta}\}$, where $\hat{\theta}$ is the estimator based on the original data.

4. The double bootstrap CI is given by $\left(\hat{\theta}^*_{\hat{q}(\frac{\alpha}{2})}, \hat{\theta}^*_{\hat{q}(1-\frac{\alpha}{2})}\right)$, where $\hat{q}(\gamma) = u^*_{(\gamma)}$, the 100γ-th percentile of the distribution of $u^{*b},\ b = 1, \ldots, B_1$.

See [31] and [63] for details about double bootstrap CIs. One disadvantage of these CIs is that they are computationally very intensive.

We use $B = 1000$ bootstrap iterations to calculate the percentile and the hybrid bootstrap CIs. However, the double bootstrap CIs are based on $B_1 = 500$ first-stage and $B_2 = 100$ second-stage bootstrap iterations (due to the increased computational burden). The results in Tables $5.2 - 5.3$ are based on $N = 1000$ Monte Carlo iterations.

5.4.1 Results

The simulation study compares the competing estimators on a variety of settings represented by Examples $1 - 6$. We considered estimation and inference for both ψ_{10} and ψ_{11}. However in the present examples, the effect of non-regularity turned out to be more pronounced for the parameter ψ_{10} (main effect of A_1) than ψ_{11} (interaction of A_1 with O_1). Hence we included results on ψ_{10} only in Tables 5.2 and 5.3. Also in the following discussion, we will focus on ψ_{10}.

In Example 1 (top part of Table 5.2), where stage 2 effects for all possible histories are zero (i.e., the stage 2 optimal treatment is non-unique for every subject in the population), we see that there is no bias associated with the hard-max estimator; and the mean squared error (MSE) is essentially the same as the variance. However the percentile bootstrap CI (both 95% and 90%) has over-coverage (note that over-coverage translates to lower power of the corresponding hypothesis test), and the hybrid bootstrap CI (95%) has under-coverage compared to the nominal level. We have also studied the Wald type CIs for this setting (not included in this paper) and observed over-coverage; the problem with Wald type CIs in such non-regular settings is well-known [71, 54]. This suggests that the asymptotic distribution of the hard-max estimator has a lighter tail than a comparable normal distribution. However, the double bootstrap CIs have correct coverage. Note that both versions of the hard-threshold estimator fail to rectify the coverage rate, even though neither suffer from bias. However, the soft-threshold estimator offers correct coverage for both types of bootstrap CIs. Moreover, it gives the lowest MSE among the four estimators. Note that the soft-threshold estimator is also non-smooth (non-regular), and consequently the bootstrap distribution is inconsistent for the true asymptotic distribution of this estimator. But in this setting, it reduces the degree of non-regularity just enough so that the bootstrap CIs do not show the problem with coverage.

Even though Example 2 (middle part of Table 5.2) is a regular setting ($p = 0$), it is very close to Example 1 and hence affected by non-regularity. Results are similar to those in Example 1. Thus the presence of very small effects causes problems with coverage even in regular settings.

Example 3 (bottom part of Table 5.2) is a setting where the stage 2 optimal treatment is non-unique for half the subjects in the population ($p = \frac{1}{2}$) and is unique for

the remaining half, but the overall standardized stage 2 effect size ϕ (= 1) is quite large. Here the hard-max estimator is biased, and hence both the percentile and the hybrid bootstrap CIs under-cover the true value. However the double bootstrap CI gives correct coverage rate. Both versions of the hard-threshold estimator reduce bias and one of them (corresponding to $\alpha = 0.08$) gives correct coverage, while the other also offers substantial improvement of the coverage rate. This is consistent with the findings of [54]. The soft-threshold estimator also reduces bias, gives the lowest MSE among the four estimators, and provides correct coverage with the hybrid bootstrap method but not with the percentile method (even though it offers substantial improvement). Thus in this example, the hard-threshold estimator with $\alpha = 0.08$ emerges as the winner, with the soft-threshold estimator at the second place. However, note that the value 0.08 of the tuning parameter α is not arbitrary – it corresponds to the threshold used by the soft-threshold estimator. If constructing

Table 5.2: Summary statistics and coverage rates of 95% and 90% nominal percentile (PB), hybrid (HB), and double (DB) bootstrap CIs for ψ_{10} using the hard-max (HM), the hard-threshold with $\alpha = 0.08$ ($HT_{0.08}$) and $\alpha = 0.2$ ($HT_{0.20}$), and the soft-threshold (ST) estimators. A "*" indicates significantly different coverage rate than the nominal rate, using a test of proportion (Type I error rate = 0.05).

	Summary Statistics			Coverage of 95% CI			Coverage of 90% CI		
Estimator	Bias	Var	MSE	PB	HB	DB	PB	HB	DB
Example 1: $p = 1$ and ϕ undefined ($\psi_{10} = \psi_{11} = 0$)									
HM	0.0003	0.0045	0.0045	96.8*	93.5*	93.6	92.9*	88.2	88.8
$HT_{0.08}$	0.0017	0.0044	0.0044	97.0*	95.0	–	93.7*	90.3	–
$HT_{0.20}$	0.0002	0.0050	0.0050	97.4*	92.8*	–	94.2*	86.9*	–
ST	0.0009	0.0036	0.0036	95.3	96.1	–	91.1	91.4	–
Example 2: $p = 0$ and ϕ infinite ($\psi_{10} = \psi_{11} = 0$)									
HM	0.0003	0.0045	0.0045	96.7*	93.4*	93.6	92.4*	88.2	89.0
$HT_{0.08}$	0.0010	0.0044	0.0044	97.1*	95.3	–	94.0*	90.5	–
$HT_{0.20}$	0.0003	0.0050	0.0050	97.3*	93.5*	–	94.3*	87.1*	–
ST	0.0008	0.0036	0.0036	95.4	95.9	–	90.8	91.5	–
Example 3: $p = \frac{1}{2}$ and $\phi = 1$ ($\psi_{10} = \psi_{11} = 0$)									
HM	-0.0401	0.0059	0.0075	88.4*	92.7*	94.8	81.2*	86.1*	89.0
$HT_{0.08}$	-0.0083	0.0058	0.0059	94.3	94.3	–	88.5	89.0	–
$HT_{0.20}$	-0.0179	0.0062	0.0065	93.5*	93.5*	–	87.0*	88.1*	–
ST	-0.0185	0.0055	0.0058	93.4*	94.9	–	87.1*	89.4	–

confidence intervals is the main goal (so biased estimation is less of an issue), double bootstrap CI along with the hard-max estimator can also be used in this setting, although it is computationally more expensive.

Example 4 (top part of Table 5.3) is a regular setting, very similar to the non-regular setting in Example 3. Results are quite similar to those in Example 3. This is consistent with our previous observation (Example 2) that the presence of very small effects causes problems with coverage even in regular settings.

In example 5 (middle part of Table 5.3), the stage 2 optimal treatment is non-unique for one-fourth of the subjects in the population ($p = \frac{1}{4}$) and the standardized effect size ϕ is very large (=1.4142). Again, the hard-max estimator is biased, and has low coverage of the CIs (except for double bootstrap). The hard-threshold and the soft-threshold estimators offer improvement in terms of bias as well as coverage. The soft-threshold estimator emerges as the best (lowest MSE and correct coverage rate) in this example.

Example 6 (bottom part of Table 5.3) is a regular setting ($p = 0$, with no extremely tiny stage 2 effect as in Examples 2 and 4), with the standardized effect size 0.3451. The reason for investigating this setting is to check if the regularized estimators (hard and soft threshold) perform poorly in settings where there is no need to regularize. As expected, the hard-max estimator performs well here. The soft-threshold estimator introduces some bias when there is none in the hard-max estimator and increases MSE; but still manages to provide correct coverage for the percentile bootstrap method. The hard-threshold estimators also give correct coverage for percentile CIs.

To summarize, the hard-max estimator is problematic in non-regular scenarios, except when used with the computationally intensive double bootstrap method for constructing confidence intervals. The hard-threshold estimator, if properly tuned,

addresses the problem of bias but not the problem of light tail. The soft-threshold estimator seems to address both problems to a large extent. In the simulation, the soft-threshold estimator consistently produced the lowest MSE among the competing methods across all the non-regular scenarios. Also in all the non-regular settings, either the soft-threshold estimator or the hard-threshold estimator with $\alpha = 0.08$ (this α corresponds to the threshold used by the soft-threshold estimator) emerged as the winner in terms of providing correct coverage rate of the bootstrap CIs. Even though the soft-threshold estimator incurs some bias in regular settings, it manages to provide reasonable coverage rate for small to moderate standardized effect sizes (we have studied up to around 0.35). Across all the scenarios considered here (Examples $1-6$), the soft-threshold estimator emerged as more robust than the hard-threshold estimator to the degree of regularity of the underlying data distribution, probably because of its "soft" nature (the soft-threshold estimator is continuous everywhere

Table 5.3: Summary statistics and coverage rates of 95% and 90% nominal percentile (PB), hybrid (HB), and double (DB) bootstrap CIs for ψ_{10} using hard-max (HM), hard-threshold with $\alpha = 0.08$ ($HT_{0.08}$) and $\alpha = 0.2$ ($HT_{0.20}$), and soft-threshold (ST) estimators. A "*" indicates significantly different coverage rate than the nominal rate, using a test of proportion (Type I error rate = 0.05).

	Summary Statistics			Coverage of 95% CI			Coverage of 90% CI		
Estimator	Bias	Var	MSE	PB	HB	DB	PB	HB	DB
Example 4: $p = 0$ and $\phi = 1.0204$ ($\psi_{10} = -0.01$, $\psi_{11} = 0$)									
HM	-0.0353	0.0059	0.0072	89.6*	93.1*	94.4	82.9*	86.6*	90.2
$HT_{0.08}$	-0.0037	0.0058	0.0058	94.6	94.1	–	88.9	89.0	–
$HT_{0.20}$	-0.0130	0.0062	0.0064	93.9	92.8*	–	87.9*	87.9*	–
ST	-0.0138	0.0055	0.0057	94.1	95.0	–	87.4*	89.7	–
Example 5: $p = \frac{1}{4}$ and $\phi = 1.4142$ ($\psi_{10} = \psi_{11} = 0$)									
HM	-0.0209	0.0069	0.0074	92.7*	93.1*	94.2	87.8*	89.0	88.4
$HT_{0.08}$	-0.0059	0.0070	0.0071	93.9	93.2*	–	89.5	88.2	–
$HT_{0.20}$	-0.0101	0.0072	0.0073	93.3*	93.0*	–	89.3	88.0*	–
ST	-0.0065	0.0069	0.0069	93.8	94.6	–	89.7	89.0	–
Example 6: $p = 0$ and $\phi = 0.3451$ ($\psi_{10} = -0.3688$, $\psi_{11} = 0.0187$)									
HM	0.0009	0.0067	0.0067	95.0	93.8	95.0	89.2	87.4*	88.2
$HT_{0.08}$	0.0003	0.0081	0.0081	95.1	88.5*	–	90.1	82.9*	–
$HT_{0.20}$	0.0011	0.0074	0.0074	94.8	91.2*	–	89.7	86.4*	–
ST	0.0052	0.0074	0.0074	94.8	91.7*	–	89.4	85.3*	–

even though it has two points of non-differentiability, whereas the hard-threshold estimator has two points of discontinuity – see Figure 5.1). Furthermore, note that overall the hybrid bootstrap CIs performed slightly better than the percentile bootstrap CIs in this simulation study. Hence the hybrid bootstrap CIs will be used in the data analysis to follow.

5.5 Analysis of the Smoking Cessation Data

To demonstrate the occurrence of non-regularity and the use of the soft-threshold method in a real application, here we present the analysis (two versions) of a data set from a randomized, two-stage, longitudinal, internet-based smoking cessation study described in chapter II. The stage 1 of this study (*Project Quit*) was conducted to find an optimal multicomponent behavioral intervention to help adult smokers quit smoking; and the stage 2 (*Forever Free*) was a follow-on study to help those (among the participants of *Project Quit*) who already quit stay quit, and help those who failed at the previous stage with a second chance. Details of the study design and primary analysis of the stage 1 data can be found in [88]; see also chapter II. In section 5.5.1 we present a complete-case analysis with the primary outcome of the study, e.g. quit status. However it is well-known that such an analysis can lead to biased estimates (so called *non-response bias*). To address this, we present a refined analysis using *multiple imputation* (MI) in section 5.5.2. See Appendix C for a brief review of the MI technique.

5.5.1 Complete-case Analysis

At stage 1, although there were five two-level treatment factors, only two, e.g., `source` (of online behavioral counseling message) and `story` (of a hypothetical character who succeeded in quitting smoking) were significant in the analysis reported in [88]. For simplicity, we considered only these two treatment factors at stage 1 of our present analysis, which gave a total of 4 treatment combinations at stage 1 corresponding to the 2×2 design. The treatment factor `source` was varied at two levels, e.g., high vs. low personalized, coded 1 and -1; also the factor `story` was varied at two levels, e.g., high vs. low tailoring depth (degree to which the character in the story was tailored to the individual subject's baseline characteristics), coded 1 and -1. We considered only a few baseline variables at this stage, e.g. `QuitOverallMotiv` (motivation to quit on a 1-10 scale), `QuitOverallSE` (self-efficacy on a 1-10 scale) and `Education` (binary, $\leq$ high school vs. $>$ high school, coded $-1/1$). At stage 2, originally there were 4 different treatment groups and a control group; however the 4 treatment groups were combined together for the present analysis because of very little difference between them. This resulted in only two choices of treatment at stage 2; this treatment variable was called `FFArm`, coded $-1/1$ (1=treatment, -1 = control).

There were two outcomes at the two stages of this study. The stage 1 outcome was binary quit status called `PQ6Quitstatus` (1=quit, 0=not quit) at 6 month from the date of randomization. The stage 2 outcome was binary quit status `FF6Quitstatus` at 6 months from the date of stage 2 randomization (i.e., 12 months from the date of stage 1 randomization).

An example DTR can have the following form: "At stage 1, if a subject's baseline self-efficacy is greater than a threshold value (say 7, on a 1-10 scale), then provide the

highly-personalized level of the treatment component `source`, and if the subject is willing to continue treatment, then at stage 2 provide treatment if s/he continues to be a smoker at the end of stage 1". Of course characteristics other than self-efficacy or a combination of more than one subject characteristics can be used to specify a DTR. To find the optimal DTR, we applied both the hard-max and the soft-threshold estimators within the Q-learning framework. This involved: (1) a stage 2 regression ($n = 281$) of `FF6Quitstatus` using the model:

$$\begin{aligned} FF6Quitstatus &= \beta_{20} + \beta_{21}\, QuitOverallMotiv + \beta_{22}\, source + \beta_{23}\, QuitOverallSE \\ &\quad + \beta_{24}\, story + \beta_{25}\, Education + \beta_{26}\, PQ6Quitstatus \\ &\quad + \beta_{27}\, source * QuitOverallSE + \beta_{28}\, story * Education \\ &\quad + \Big(\psi_{20} + \psi_{21}\, PQ6Quitstatus\Big) * FFArm + \epsilon_2; \end{aligned}$$

(2) finding both the hard-max pseudo-outcome ($\hat{Y}_1$) and the soft-threshold pseudo-outcome ($\hat{Y}_1^{ST}$) for the stage 1 regression:

$$\begin{aligned} \hat{Y}_1 &= PQ6Quitstatus + \hat{\beta}_{20} + \hat{\beta}_{21}\, QuitOverallMotiv + \hat{\beta}_{22}\, source \\ &\quad + \hat{\beta}_{23}\, QuitOverallSE + \hat{\beta}_{24}\, story + \hat{\beta}_{25}\, Education + \hat{\beta}_{26}\, PQ6Quitstatus \\ &\quad + \hat{\beta}_{27}\, source * QuitOverallSE + \hat{\beta}_{28}\, story * Education \\ &\quad + \Big|\hat{\psi}_{20} + \hat{\psi}_{21}\, PQ6Quitstatus\Big|; \end{aligned}$$

$$\begin{aligned} \hat{Y}_1^{ST} &= PQ6Quitstatus + \hat{\beta}_{20} + \hat{\beta}_{21}\, QuitOverallMotiv + \hat{\beta}_{22}\, source \\ &\quad + \hat{\beta}_{23}\, QuitOverallSE + \hat{\beta}_{24}\, story + \hat{\beta}_{25}\, Education + \hat{\beta}_{26}\, PQ6Quitstatus \\ &\quad + \hat{\beta}_{27}\, source * QuitOverallSE + \hat{\beta}_{28}\, story * Education \\ &\quad + \Big|\hat{\psi}_{20} + \hat{\psi}_{21}\, PQ6Quitstatus\Big| \cdot \left(1 - \frac{3\mathrm{Var}(\hat{\psi}_{20} + \hat{\psi}_{21}\, PQ6Quitstatus)}{|\hat{\psi}_{20} + \hat{\psi}_{21}\, PQ6Quitstatus|^2}\right)^{+}; \end{aligned}$$

and (3) for each of the two pseudo-outcomes, a stage 1 regression ($n = 1401$) of the

pseudo-outcome using a model of the form:

$$\begin{aligned}\hat{Y}_1 \text{ or } \hat{Y}_1^{ST} &= \beta_{10} + \beta_{11}\, QuitOverallMotiv + \beta_{12}\, QuitOverallSE + \beta_{13}\, Education \\ &\quad + \left(\psi_{10}^{(1)} + \psi_{11}^{(1)}\, QuitOverallSE\right) * source \\ &\quad + \left(\psi_{10}^{(2)} + \psi_{11}^{(2)}\, Education\right) * story + \epsilon_1.\end{aligned}$$

Note that the sample sizes at the two stages differ because only 281 subjects were willing to continue treatment into stage 2 (as allowed by the study protocol). Our stage 2 analysis was a usual regression analysis. No significant treatment effect was found at this stage, indicating the likely existence of non-regularity. At stage 1, for either estimator, 95% confidence intervals were constructed by hybrid bootstrap using 1000 bootstrap replications. The stage 1 analysis summary is presented in Table 5.4. In this case, the hard-max and the soft-threshold estimators produced similar results.

Table 5.4: Regression coefficients and 95% hybrid bootstrap confidence intervals at stage 1, using both the hard-max and the soft-threshold estimators.

	Hard-max		Soft-threshold	
Variable	Coefficient	95% CI	Coefficient	95% CI
`QuitOverallMotiv`	0.04	(-0.00, 0.08)	0.04	(0.00, 0.08)
`QuitOverallSE`	0.03	(0.00, 0.06)	0.03	(0.00, 0.06)
`Education`	-0.01	(-0.07, 0.06)	-0.01	(-0.07, 0.06)
`source`	-0.15	(-0.35, 0.06)	-0.15	(-0.35, 0.06)
`source*QuitOverallSE`	0.03	(0.00, 0.06)	0.03	(0.00, 0.06)
`story`	0.05	(-0.01, 0.11)	0.05	(-0.01, 0.11)
`story*Education`	-0.07	(-0.13, -0.01)	-0.07	(-0.13, -0.01)

The conclusions from the present data analysis can be summarized as follows. We did not find any significant stage 2 treatment effect. So this analysis suggests that the stage 2 behavioral intervention need not be adapted to the smoker's individual characteristics, interventions previously received, or stage 1 outcome. More interesting results are found at stage 1. It is found that subjects with higher level of motivation or self-efficacy are more likely to quit. The highly personalized level

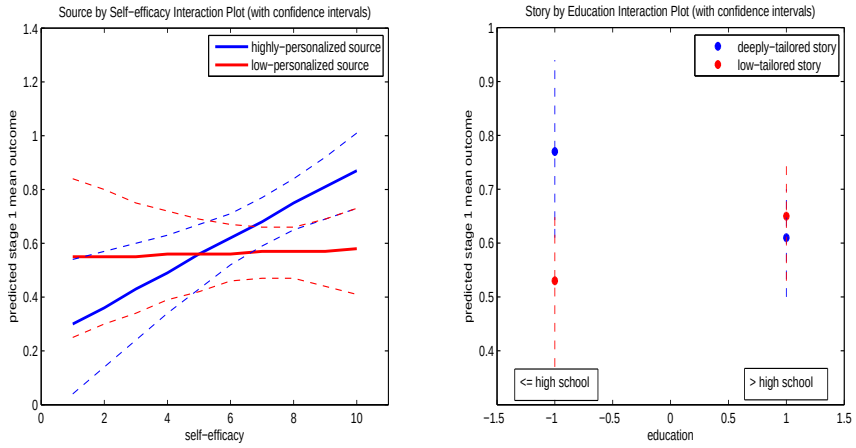

Figure 5.2: Interaction plots: (a) source by self-efficacy (left panel), (b) story by education (right panel), along with confidence intervals for predicted stage 1 pseudo-outcome.

of `source` is more effective for subjects with a higher self-efficacy (≥ 7), and deeply tailored level of `story` is more effective for subjects with lower education ($\leq$ high school); these two conclusions can be drawn from the interaction plots (with confidence intervals) presented in figure 5.2. Thus this analysis suggests that to maximize each individual's chance of quitting over the two stages, the web-based smoking cessation intervention should be designed in future such that: (1) smokers with high self-efficacy (≥ 7) are assigned to highly personalized level of `source`, and (2) smokers with lower education are assigned to deeply tailored level of `story`.

5.5.2 Analysis using Multiple Imputation

As mentioned in chapter II, the study started with 1848 subjects (Project Quit), out of which 479 consented to move to stage 2 (Forever Free). Among these, there were 1401 complete cases at stage 1 and 281 complete cases in stage 2, and were included in the complete-case analysis presented above. The completely observed variables (X) in the study were: `HMO` (binary), `gender` (binary), `age` (continuous),

`QuitCigsPerDay` (baseline number of cigarettes smoked per day, continuous), and the five stage-1 treatment components, e.g., `source` (personalization of source), `outcome.depth` (tailoring depth of outcome expectations), `story` (tailoring depth of success stories), `efficacy.depth` (tailoring depth of efficacy expectations), and `exposure` – each with 2 levels. The remaining variables (Y) along with their missingness rates are listed below.

Table 5.5: Baseline Variables subject to Missingness

Variables	Number of Missing Values	Rate of Missingness (%)
`QuitOverallMotiv` (motivation, 1-10, continuous)	5	0.27
`QuitOverallSE` (self-efficacy, 1-10, continuous)	4	0.21
`Education` (binary, high school vs. college)	7	0.38
`RaceWhite` (binary, 1 if the subject is White)	10	0.54
`RaceBlack` (binary, 1 if the subject is Black)	10	0.54

The variables `RaceWhite` and `RaceBlack` are two dummy variables constructed from the original three-level variable `Race`. The rates of missingness are very small in the baseline variables. Below we list the missingness rates of the stage-1 variables.

Table 5.6: Stage-1 (collected at 6 months) Variables subject to Missingness

Variables	Number of Missing Values	Rate of Missingness (%)
`PQ6Quitstatus` (binary, 1= quitter, 0=smoker)	436	23.59
`PQ6MonthsNS` (months not smoked, 0-6)	699	37.82
`PQ6OverallSat` (overall satisfaction, 1-10)	439	23.76
`PQ6NumOfAttempts` (number of quit attempts)	656	35.50
`PQ6OverallMotiv` (motivation, 1-10)	440	23.81
`PQ6OverallSE` (self-efficacy, 1-10)	442	23.92

Next we discuss our strategy to model different variables for the purpose of multiple imputation. The variables `PQ6OverallMotiv` (and `QuitOverallMotiv`), `PQ6OverallSE` (and `QuitOverallSE`) and `PQ6OverallSat` are all left-skewed and vary in the range $1 - 10$; for convenience of modeling we binarize them (cutting off at the respective means leading to two levels, high vs. low). The corresponding

binarized variables go by the same name as the original variable, except for the suffix `Bin`. Histograms of the remaining stage-1 variables are presented below.

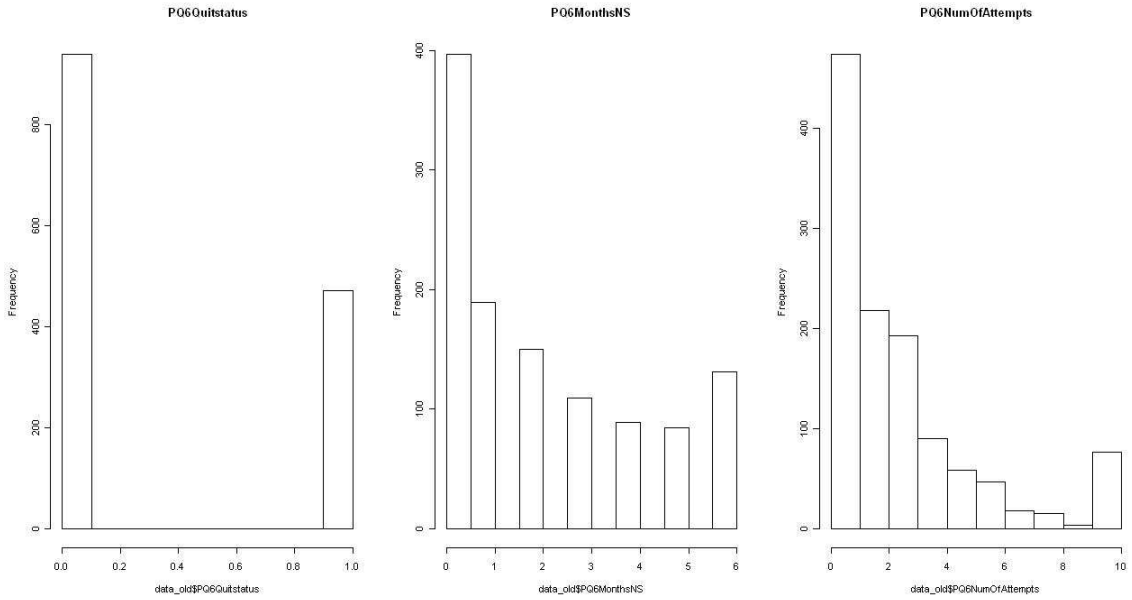

Figure 5.3: Histograms of PQ6Quitstatus, PQ6MonthsNS, and PQ6NumOfAttempts.

From the histograms above, we see that:

1. `PQ6MonthsNS` cannot be modeled by a normal distribution; a better modeling strategy would be to use 6 dummy variables to represent it, each of which will be approximated by a normal and then rounded off.

2. `PQ6NumOfAttempts` is right-skewed; a square-root transformation may be better (see below).

Next, let us look at the stage-2 variables. Among the 479 subjects who moved to stage 2 (Forever Free), the stage-2 treatment, `FFArm` (2 levels) is completely observed. Below we list the missingness rates (out of 479 subjects) of the stage-2 variables.

Clearly, the rates of missingness are pretty high at stage 2. As in stage 1, we binarize the variable denoting the level of satisfaction with the smoking cessation

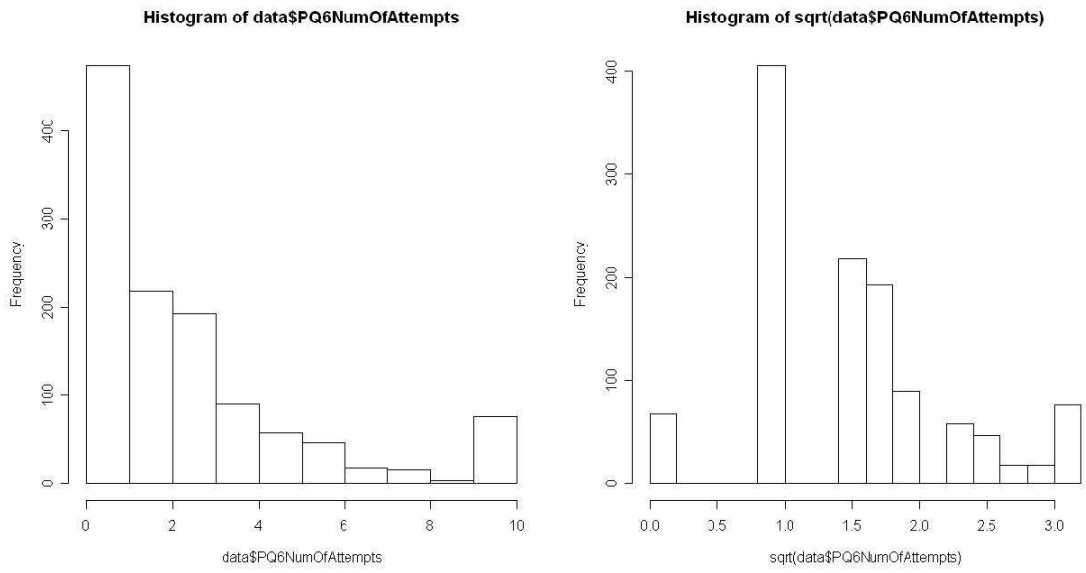

Figure 5.4: Histograms of PQ6NumOfAttempts and its square-root.

Table 5.7: Stage-2 (collected at 12 months) Variables subject to Missingness

Variables	Number of Missing Values	Rate of Missingness (%)
`FF6Quitstatus` (binary, 1= quitter, 0=smoker)	195	40.71
`FF6MonthsNS` (months not smoked, 0-6)	264	55.11
`FF6OverallSat` (overall satisfaction, 1-10)	205	42.80
`FF6NumOfAttempts` (number of quit attempts)	263	54.91

program (`FF6OverallSat`). Let us look at the histograms of the remaining stage-2 variables. From the histograms, we see that:

1. `FF6MonthsNS` clearly cannot be modeled by a normal distribution. As in stage 1, the first alternative strategy would be to use 6 dummy variables to represent the original variable, and approximate each dummy variable by a normal distribution and then round off. However, note that the frequency corresponding to some of the levels (e.g. 3, 4, 5) are really small, making the success probabilities of the corresponding dummies very small – a situation where normal approximation fails to do a good job [4]. To address this issue, the variable is

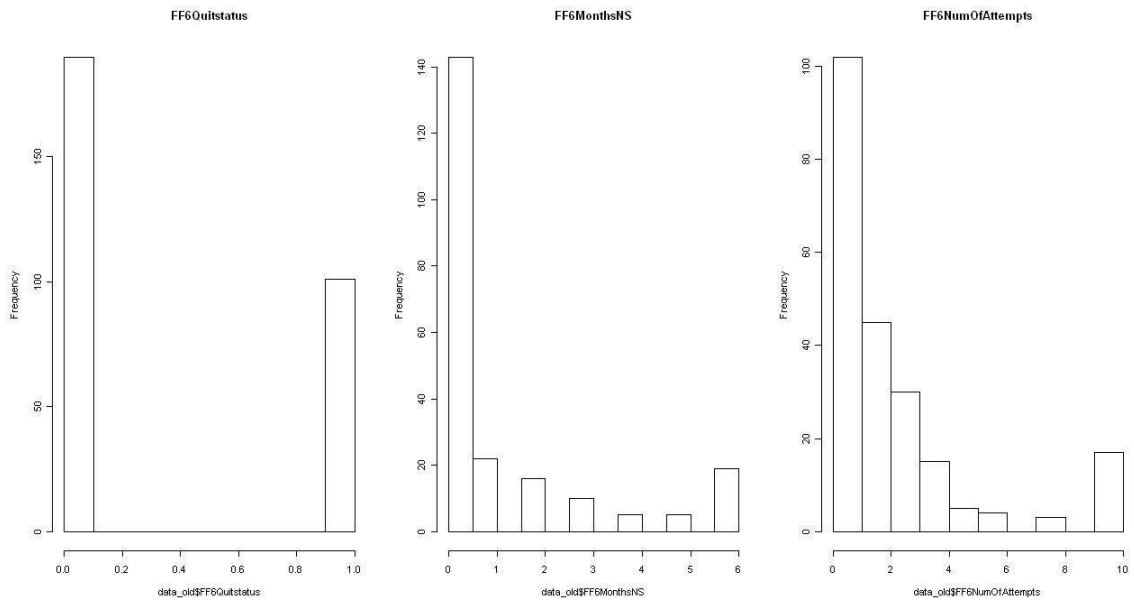

Figure 5.5: Histograms of FF6Quitstatus, FF6MonthsNS, and FF6NumOfAttempts.

re-coded with only three levels, 0, 1, and 2: it takes the value 0 if the original `FF6MonthsNS` is 0, takes the value 1 if 0<`FF6MonthsNS`<=3, and takes the value 2 if 3<`FF6MonthsNS`<=6. For consistency among the stages, the corresponding stage-1 variable, `PQ6MonthsNS` is also re-coded accordingly.

2. `FF6NumOfAttempts` is right-skewed; a square-root transformation may be somewhat better although not completely satisfactory (see below).

Next let us examine the histograms of the newly constructed outcome variables (trinary version of `PQ6MonthsNS` and `FF6MonthsNS`).

Since `PQ6MonthsNS` is sort of symmetric and unimodal (and of course ordinal), instead of breaking it into two dummies, it is directly modeled by a normal distribution and then rounded off. On the other hand, `FF6MonthsNS` is modeled using two dummies, using a normal approximation for each.

We used multiple imputation under a multivariate normal model (see Appendix C). The MCMC procedure for MI converged nicely with the default non-informative

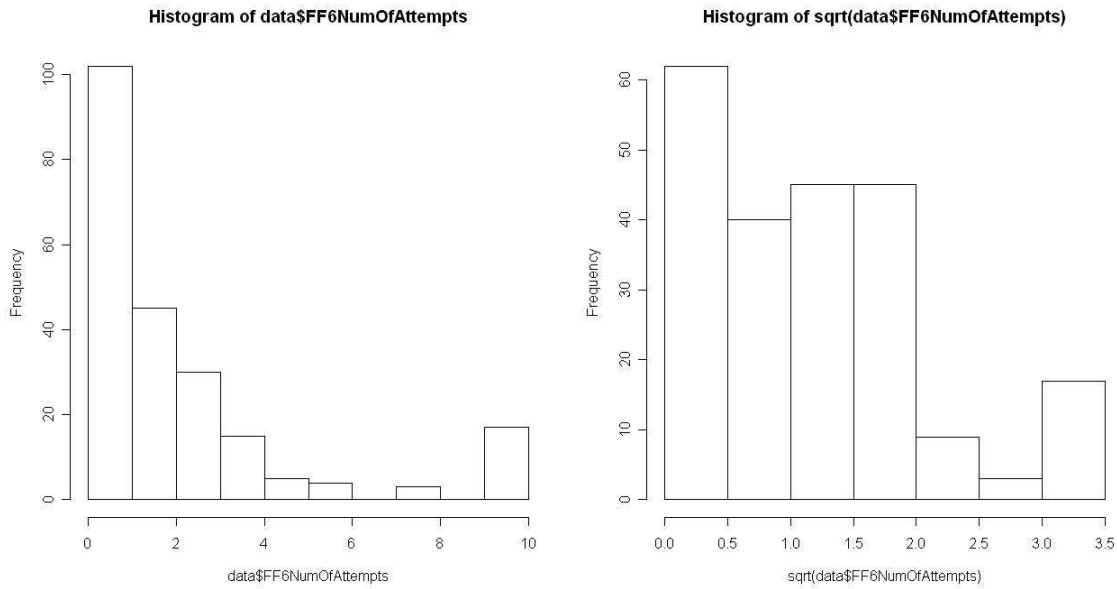

Figure 5.6: Histograms of FF6NumOfAttempts and its square-root.

priors. A burn-in period of 1000 iterations was used in the MCMC procedure. The following graph provides a global check of convergence of the MCMC – the auto-correlation function of the "worst cosine function" as described by Schafer [80]. Even

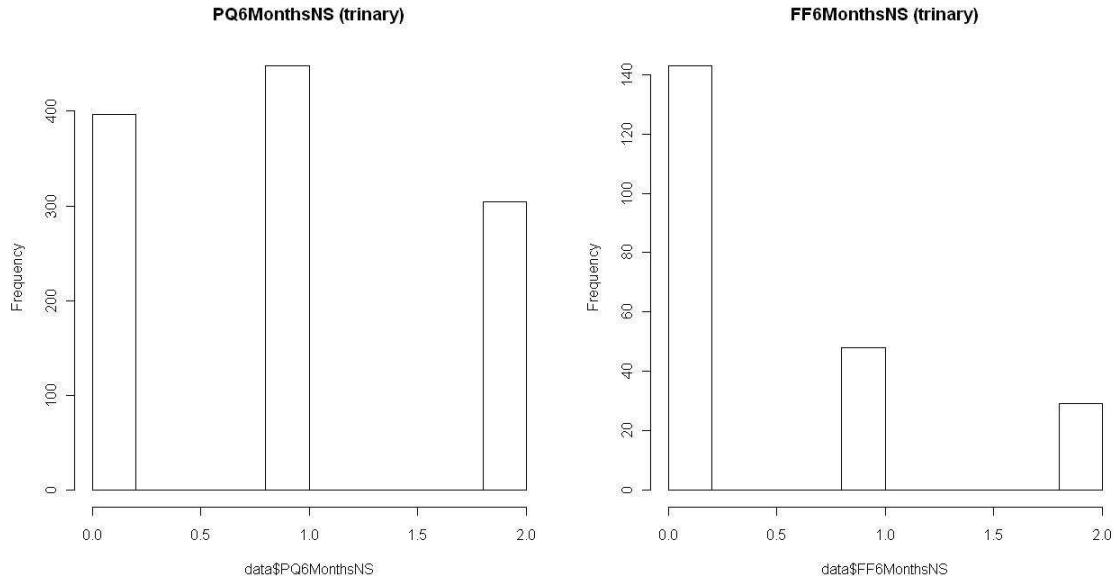

Figure 5.7: Histograms of PQ6MonthsNS and FF6MonthsNS (trinary).

though in most cases the convergence of the above global criterion ensures convergence of the individual parameters, there is no theoretical guarantee (and there exist counterexamples). None-the-less, it is a good starting point to assess the convergence of the MCMC procedure.

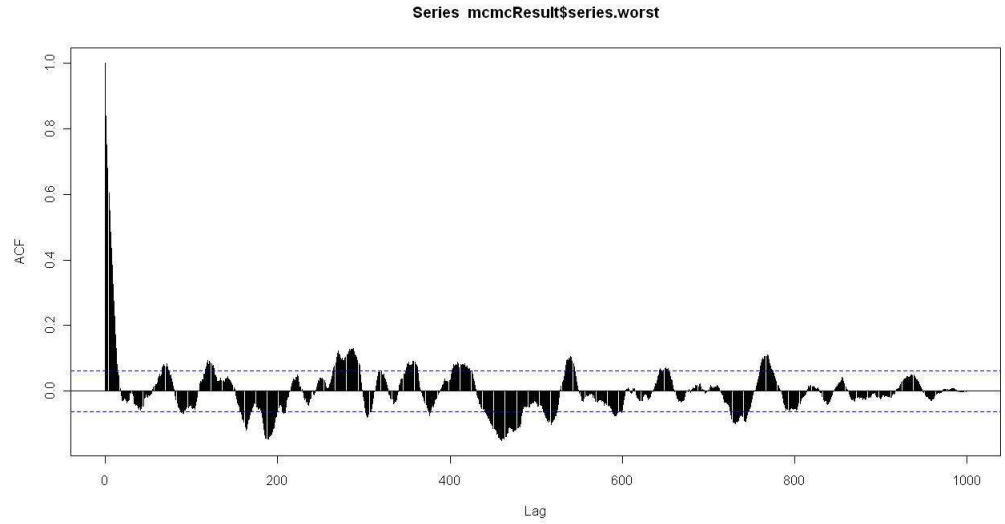

Figure 5.8: Auto-correlation function of the worst cosine function.

Next we present some checks for the individual parameters of interest, e.g. the regression parameters associated with the variables `PQ6MonthsNS`, `FF6MonthsNSDummy1` and `FF6MonthsNSDummy2`, since these will be considered as the primary outcome variables in the data analysis to follow.

Next we present the actual analysis of the multiply imputed data set. The point estimates (regression coefficients in the Q-learning analysis) reported here are averages of the corresponding coefficients over 10 imputed data sets. Measures of confidence are provided by bootstrap CIs. We use 500 bootstrap samples, and 10 imputations within each bootstrap sample. Within a given bootstrap sample, the regression coefficients in the model for Q-functions are created by averaging over 10 imputed datasets. Finally, the quantiles of these bootstrapped coefficients (each one

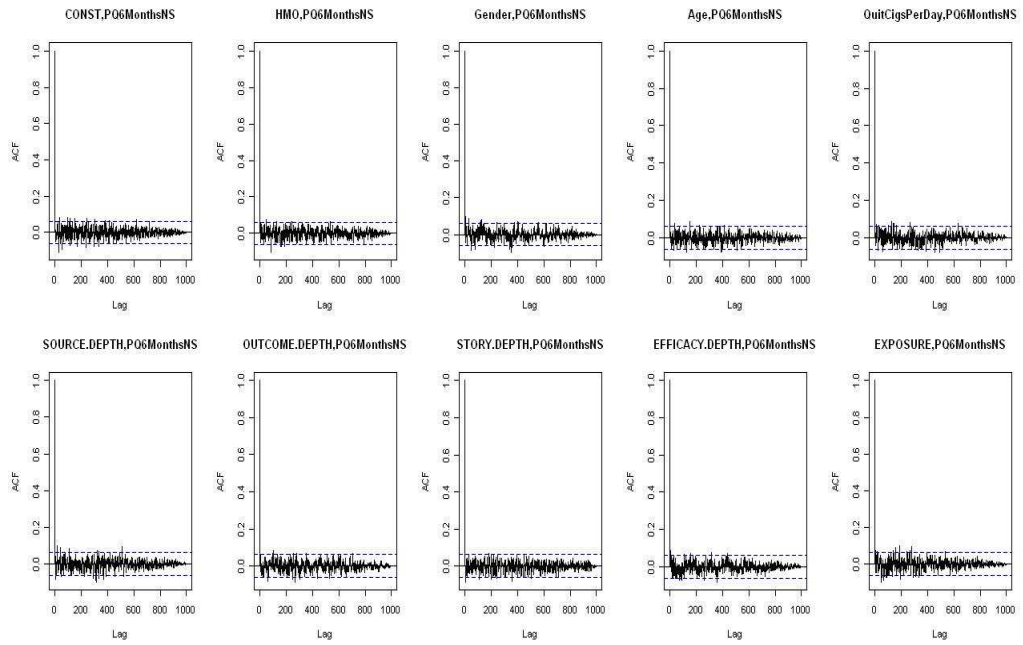

Figure 5.9: Auto-correlation functions of the regression parameters associated with the stage-1 outcome (PQ6MonthsNS).

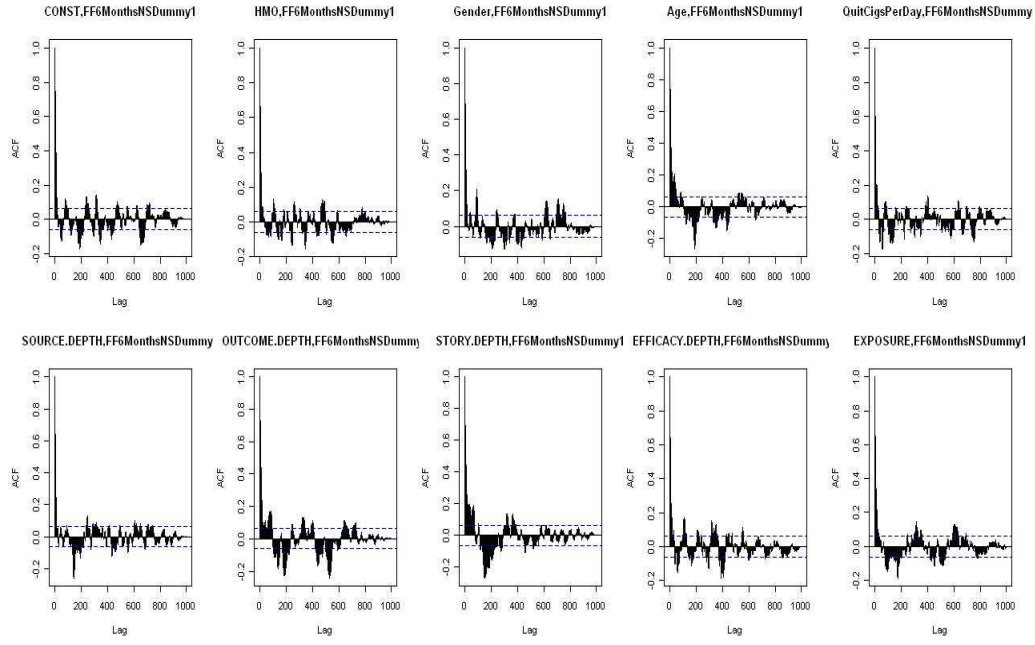

Figure 5.10: Auto-correlation functions of the regression parameters associated with the first dummy variable corresponding to the stage-2 outcome (FF6MonthsNSDummy1).

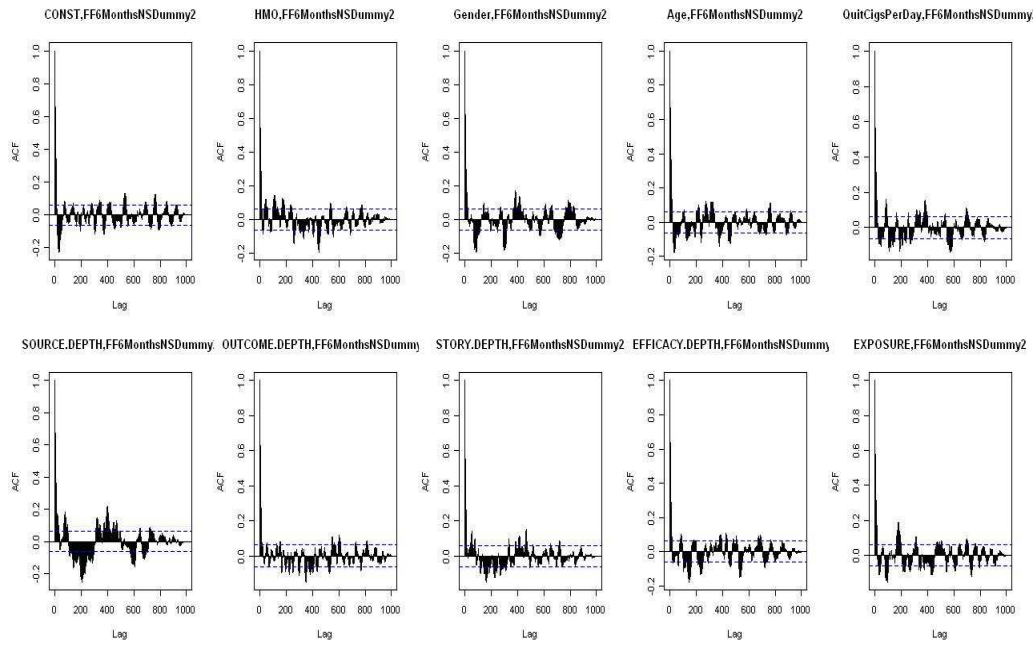

Figure 5.11: Auto-correlation functions of the regression parameters associated with the second dummy variable corresponding to the stage-2 outcome (FF6MonthsNSDummy2).

is an average over 10 imputations) are used to create the hybrid bootstrap CIs.

Here we use a very parsimonious analysis model, with `story` as the only stage-1 treatment variable, `Education` as the only stage-1 covariate, and `PQ6Quitstatus` as the only stage-2 covariate. The outcome variables are `PQ6MonthsNS` (stage 1) and `FF6MonthsNS` (stage 2), both taking values 0, 1, 2. The stage-2 regression model ($n = 479$) is given by:

$$\begin{aligned} FF6MonthsNS \quad \sim \quad & story + Education + PQ6Quitstatus + \\ & story * Education + FFArm + PQ6Quitstatus * FFArm \end{aligned}$$

Based on the estimated coefficients of the stage-2 regression, both the hard-max and the soft-threshold pseudo-outcomes are constructed (as in section 5.5.1). Then the stage-1 regression model ($n = 1848$) is given by:

$$Pseudooutcome \sim story + Education + story * Education$$

No significant stage-2 treatment effect was found (consistent with the complete case analysis), indicating the likely existence of non-regularity. Results of the stage-1 regression analysis is presented below.

Table 5.8: Regression coefficients and 95% hybrid bootstrap confidence intervals at stage 1, using both the hard-max and the soft-threshold estimators, and using multiple imputation.

	Hard-max		Soft-threshold	
Variable	Coefficient	95% CI	Coefficient	95% CI
`Education`	0.011	(-0.180, 0.199)	0.012	(-0.178, 0.200)
`story`	0.072	(-0.074, 0.195)	0.070	(-0.076, 0.192)
`story*Education`	-0.112	(-0.240, -0.004)*	-0.111	(-0.237 -0.002)*

Next we present the lengths of the bootstrap CIs for the two competing methods of estimation; the soft-threshold method looks marginally better.

Table 5.9: Length of Bootstrap CIs.

Variable	Hard-max	Soft-threshold
Education	0.379	0.378
story	0.269	0.268
story:Education	0.236	0.235

In Table 5.8, a negative interaction between `story` and `Education` is detected. This is one of the interactions anticipated a priori by the study investigators; see [88]. This finding is also consistent with the complete case analysis in section 5.5.1. This negative interaction is interpretable – it says that highly individually tailored level of `story` (success story of a hypothetical smoker) is significantly more effective within smokers with low education (high school graduate or less, but no college education). This finding allows one to individually tailor the smoking cessation intervention.

5.6 Discussion

In this chapter, we have illustrated the problem of non-regularity that arises in the context of DTRs in the estimation of the optimal current treatment rule, when the

optimal treatments at subsequent stages are non-unique for at least some proportion of subjects in the population. We have illustrated the phenomenon using Q-learning as the estimation procedure, which is a simpler yet inefficient version of Robins' method; however the problem of non-regularity arises in Robins' method as well [71, 54].

For some underlying data-generating models (e.g., Examples 3, 4, 5 in the simulation study), this non-regularity induces bias in the point estimates of the parameters of the optimal DTRs, which in turn causes under-coverage of the bootstrap confidence intervals. In contrast, in case of Examples 1 and 2, this non-regularity causes lightness of tail of the asymptotic distribution but no bias, as seen from the over-coverage of the percentile bootstrap CIs (equivalently conservative tests leading to lower power). The coexistence of these two not-so-well-related issues (they work in opposite directions, e.g., bias tends to make the CIs under-cover, whereas lightness of tail tends to make the CIs over-cover) makes this problem a unique and challenging one.

In the simulation study to compare the competing estimators of the optimal DTR, we considered estimation of ψ_{10}, which involve linear combinations of $|f_1|$, $|f_2|$, $|f_3|$, and $|f_4|$ (terms like $|\mu|$). Under the non-regular scenarios, some or all (depending on the degree of non-regularity p) of the f_i's are zero; and hence a phenomenon similar to the one described above in the toy example happens for each $|f_i|$ for which $f_i = 0$. Each such term has its associated bias, and each has its own lightness of tail, with bias being the dominant property. In some non-regular scenarios (Example 1), the bias associated with the individual $|f_i|$'s (in the expression for ψ_{10}) cancel each other out (note the opposite signs in front of $|f_i|$'s), and hence the lightness of tail is revealed, resulting in a percentile bootstrap CI that over-covers. In other non-regular

examples, however, bias is not canceled out, and hence dominates the property of the hard-max estimator. Hence under-coverage of the bootstrap CIs is observed.

Non-regularity is an issue in the estimation of the optimal DTRs because it arises when there is no treatment effect at subsequent stages (equivalently, there is no unique optimal treatment at subsequent stages). Unfortunately often there is no or very weak treatment effect in the settings we are interested in (e.g., randomized trials on mental illness or substance abuse). Thus we want our estimator to enjoy good statistical properties (e.g., less bias, lower risk or MSE, correct coverage rate of CIs, good power to detect "local" alternatives, etc.) when the optimal treatment at subsequent stages is non-unique. In case of the hard-max estimator, unfortunately the point of non-differentiability coincides with the parameter value such that $\psi_2^T H_{21} = 0$ (non-unique optimal treatment at the subsequent stage), which causes non-regularity (bias, higher MSE, low power). But the soft-threshold estimator (also, hard-threshold estimator), in some sense, redistributes the non-regularity from this "null point" to two different points symmetrically placed on either side of the "null point" (see Figure 5.1). This is one reason why the soft-threshold (also, hard-threshold) estimator works well in non-regular settings.

We have shown that using bootstrap confidence intervals along with the soft-threshold (also, hard-threshold in some cases) estimator reduces the degree of non-regularity, and gives correct coverage rate. Also, the double bootstrap method can be used along with the original hard-max estimator to address the non-regularity. But this method is highly computationally intensive and may be difficult to use in practice. An alternative method to construct CIs for ψ's in non-regular settings is the score method due to Robins [71]. We have not investigated this in our simulation study.

One can consider an alternative Bayesian approach to formulate an estimator similar to the soft-threshold estimator as follows. Let the data distribution $\hat{\psi}_2^T H_{21} \mid \psi_2^T H_{21} \sim N(\psi_2^T H_{21}, \sigma^2)$ with known σ^2, and the prior distribution of $\psi_2^T H_{21}$ be a mixture of a point mass at 0 and $N(0,1)$, with mixing parameter p $(0 < p < 1)$. Then the posterior distribution of $\psi_2^T H_{21}$ is a mixture distribution given by

$$\begin{aligned} f_{post}(\psi_2^T H_{21}) &= \hat{w} \cdot \mathbf{1}\{\psi_2^T H_{21} = 0\} + (1 - \hat{w}) \cdot N\Big(\frac{\hat{\psi}_2^T H_{21}}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\Big), \\ \text{where} \quad \hat{w} &= \left\{1 + \frac{1-p}{p}\sqrt{\frac{\sigma^2}{1+\sigma^2}}\exp\left\{\frac{(\hat{\psi}_2^T H_{21})^2}{2\sigma^2(1+\sigma^2)}\right\}\right\}^{-1}. \end{aligned}$$

One can use the median of this posterior distribution in place of $\hat{\psi}_2^T H_{21}$ in the expression for $\hat{Y}_1$. Thus the Bayes estimator becomes

$$\hat{Y}_1^{Bayes} = \hat{\beta}_2^T H_{20} + \text{median of } f_{post}(\psi_2^T H_{21}).$$

For using this, one has to replace σ^2 by $\hat{\sigma}^2 = H_{21}^T \hat{\Sigma}_2 H_{21}/n$, and p by either some empirical estimate or a fixed value (e.g., $\frac{1}{2}$). Johnstone and Silverman [46] suggest using the mixture of a point mass and a heavy-tailed distribution (e.g., double-exponential) in place of the above mixture prior. This is a formulation that we want to investigate in future. Also, fully Bayesian approaches to handle the problem of estimating DTRs demand serious attention; we will consider this in future.

In this chapter, we have focused on randomized trials only to separate the issue of non-regularity from causal inference issues. However the problem of non-regularity also arises when observational data are used [71, 54]; and the hard-threshold and the soft-threshold estimators should be applicable in those settings as well. Also, here we have focussed on only two stages for clarity. However, it should be understood that Q-learning can be used for studies with more than two stages as well. In case of many stages, one can think of a scenario where some parameters are shared across

stages, in which case a simultaneous version of Q-learning (as opposed to the recursive version discussed in this chapter) would be more appropriate. Unfortunately non-regularity does not go away if a simultaneous estimation procedure is used; see [54] for a discussion on this with reference to Robins' method. However, unlike the case of recursive estimation, it is not well understood at this point whether the threshold estimators (hard or soft) can reduce the non-regularity in simultaneous estimation. Moodie and Richardson [54] gave a simulated non-regular example showing that hard-threshold or ZIPI estimator is not always better than simultaneous estimator of Robins. We did not investigate this issue here, but we recognize this as an important avenue of future research.

To conclude, we think in the estimation of optimal DTRs, appropriately tuned hard-threshold estimator and the soft-threshold estimator should be seriously considered as improved versions of Q-learning (and Robins' method of estimation).

5.7 Appendix A: Proof of Lemma V.1

Proof. Define the *advantage* at stage j as

$$\mu_j(H_j, A_j) = Q_j(H_j, A_j) - \max_{a_j} Q_j(H_j, a_j),\ j = 1, 2.$$

Note that $\mu_j(H_j, A_j)$ represents the expected difference in outcome when using A_j instead of the optimal treatment at stage j, for subjects with treatment and covariate history H_j who receive the optimal DTR at stages subsequent to j. According to Robins [71, p. 201], this is simply the *blip* function with $\arg\max_{a_j} Q_j(H_j, a_j)$ as the reference treatment. Below we will establish the connection between Q-learning and Robins' method using the advantage function; one can derive the connection using other blip functions (other choices of reference treatment) following similar steps.

When Q-functions are modeled as in (5.2), the advantages become

$$\mu_j(H_j, A_j; \psi_j) = \psi_j^T H_{j1} A_j - |\psi_j^T H_{j1}|,\ j = 1, 2. \tag{5.11}$$

Since by condition (i), no parameters are shared across stages, we will proceed stage by stage, starting with stage 2, doing recursive (rather than simultaneous) estimation. The notation $\mathbb{P}_n$ will be used below to denote the empirical average over a sample of size n. Also, define $m_1(H_1) = E[Q_1(H_1, A_1)|H_1]$ and $m_2(H_2) = E[Q_2(H_2, A_2)|H_2]$.

Stage 2:

At stage 2, Q-learning is a usual least squares regression problem. Thus the estimating equations are given by

$$\mathbb{P}_n\Big[\begin{pmatrix} H_{20} \\ H_{21}A_2 \end{pmatrix} (Y_2 - H_{20}^T\beta_2 - H_{21}^T A_2\psi_2)\Big] = 0. \tag{5.12}$$

From (5.12), it follows that

$$\hat{\beta}_2 = (\mathbb{P}_n(H_{20}H_{20}^T))^{-1}\Big[\mathbb{P}_n(H_{20}Y_2) - \mathbb{P}_n(H_{20}H_{21}^T A_2)\hat{\psi}_2\Big] \tag{5.13}$$

where $\hat{\psi}_2$ is the estimate of ψ_2 satisfying (5.12). Thus $\hat{\psi}_2$ satisfies the estimating equation

$$\mathbb{P}_n\Big[(H_{21}A_2)(Y_2 - H_{20}^T\hat{\beta}_2 - H_{21}^T A_2\hat{\psi}_2)\Big] = 0.$$

On the other hand, the stage 2 estimating equation for Robins' method [71, p. 211] is given by

$$\mathbb{P}_n\Bigg[\Big(H_{21}A_2 - E[H_{21}A_2|H_2]\Big)\Big(Y_2 - \mu_2(H_2, A_2; \psi_2) - E[Y_2 - \mu_2(H_2, A_2; \psi_2)|H_2]\Big)\Bigg] = 0, \tag{5.14}$$

where $Var\Big(Y_2-\mu_2(H_2,A_2;\psi_2)-E[Y_2-\mu_2(H_2,A_2;\psi_2)|H_2]\Big|H_2,A_2\Big)$ is omitted (This is one of the reasons why Q-learning is an inefficient version). Note that $E[H_{21}A_2|H_2]=0$, by condition (ii) of the lemma. From (5.11), $\mu_2(H_2,A_2;\psi_2)=\psi_2^TH_{21}A_2-|\psi_2^TH_{21}|$. Then $E[\mu_2(H_2,A_2;\psi_2)|H_2]=-|\psi_2^TH_{21}|$, again by condition (ii). Also,

$$E[Y_2|H_2]=E\Big[E[Y_2|H_2,A_2]|H_2\Big]=E[Q_2(H_2,A_2)|H_2]=m_2(H_2).$$

Therefore, $Y_2-\mu_2(H_2,A_2;\psi_2)-E[Y_2-\mu_2(H_2,A_2;\psi_2)|H_2]=Y_2-m_2(H_2)-H_{21}^TA_2\psi_2$. Thus, $\hat{\psi}_2$ in Robins' method solves the following reduced version of (5.14):

$$\mathbb{P}_n\Bigg[\Big(H_{21}A_2\Big)\Big(Y_2-m_2(H_2)-H_{21}^TA_2\hat{\psi}_2)\Bigg]=0,$$

for any choice of $m_2(H_2)$ (with the conditional variance omitted). In particular, for $m_2(H_2)=H_{20}^T\hat{\beta}_2$, where $\hat{\beta}_2$ is given by (5.13), this estimating equation exactly matches with that of Q-learning.

Stage 1:

For Q-learning, the stage 1 pseudo-outcome is

$$\hat{Y}_1=Y_1+\max_{a_2}Q_2(H_2,A_2)=Y_1+H_{20}^T\hat{\beta}_2+|\hat{\psi}_2^TH_{21}|,$$

and so the estimating equations are given by

$$(5.15)\ \mathbb{P}_n\Big[\begin{pmatrix}H_{10}\\H_{11}A_1\end{pmatrix}\Big(Y_1+H_{20}^T\hat{\beta}_2+|\hat{\psi}_2^TH_{21}|-H_{10}^T\beta_1-H_{11}^TA_1\psi_1\Big)\Big]=0.$$

Now from (5.12),

$$(5.16)\qquad \mathbb{P}_n\Big[H_{20}\Big(Y_2-H_{20}^T\hat{\beta}_2-H_{21}^TA_2\hat{\psi}_2\Big)\Big]=0.$$

Since by condition (iii) of the lemma, $(H_{10}^T, H_{11}^T A_1) \subset H_{20}^T$, it follows that

$$\mathbb{P}_n\Big[\begin{pmatrix} H_{10} \\ H_{11}A_1 \end{pmatrix}\Big(Y_2 - H_{20}^T\hat{\beta}_2 - H_{21}^T A_2\hat{\psi}_2\Big)\Big] = 0,$$

(5.17) or, $\mathbb{P}_n\Big[\begin{pmatrix} H_{10} \\ H_{11}A_1 \end{pmatrix}(H_{20}^T\hat{\beta}_2)\Big] = \mathbb{P}_n\Big[\begin{pmatrix} H_{10} \\ H_{11}A_1 \end{pmatrix}(Y_2 - H_{21}^T A_2\hat{\psi}_2)\Big].$

Using (5.17) in (5.15), we get

$$\mathbb{P}_n\Big[\begin{pmatrix} H_{10} \\ H_{11}A_1 \end{pmatrix}\Big(Y_1 + Y_2 - H_{21}^T A_2\hat{\psi}_2 + |\hat{\psi}_2^T H_{21}| - H_{10}^T\beta_1 - H_{11}^T A_1\psi_1\Big)\Big] = 0.$$

(5.18)

Solving for β_1 gives,

$$\hat{\beta}_1 = (\mathbb{P}_n(H_{10}H_{10}^T))^{-1}\Big[\mathbb{P}_n\Big(H_{10}(Y_1 + Y_2 - H_{21}^T A_2\hat{\psi}_2 + |\hat{\psi}_2^T H_{21}|)\Big) - \mathbb{P}_n(H_{10}H_{11}^T A_1)\hat{\psi}_1\Big].$$

(5.19)

Thus $\hat{\psi}_1$ satisfies

$$\mathbb{P}_n\Big[(H_{11}A_1)\Big(Y_1 + Y_2 - H_{21}^T A_2\hat{\psi}_2 + |\hat{\psi}_2^T H_{21}| - H_{10}^T\hat{\beta}_1 - H_{11}^T A_1\hat{\psi}_1\Big)\Big] = 0.$$

On the other hand for Robins' method, the stage 1 pseudo-outcome (e.g. [71, p. 208]; see also [54]) is $\tilde{Y}_1 = Y_1 + Y_2 - \mu_2(H_2, A_2)$, and so the stage 1 estimating equation [71, p. 211] is given by

$$\mathbb{P}_n\Bigg[\Big(H_{11}A_1 - E[H_{11}A_1|H_1]\Big)\Big(\tilde{Y}_1 - \mu_1(H_1, A_1; \psi_1) - E[\tilde{Y}_1 - \mu_1(H_1, A_1; \psi_1)|H_1]\Big)\Bigg] = 0,$$

(5.20)

where again the conditional variance

$$Var\Big(\tilde{Y}_1 - \mu_1(H_1, A_1; \psi_1) - E[\tilde{Y}_1 - \mu_1(H_1, A_1; \psi_1)|H_1]\Big|H_1, A_1\Big)$$

is omitted. Note that $E[H_{11}A_1|H_1] = 0$, by condition (ii) of the lemma. From (5.11), $\mu_1(H_1, A_1; \psi_1) = \psi_1^T H_{11} A_1 - |\psi_1^T H_{11}|$. Then $E[\mu_1(H_1, A_1; \psi_1)|H_1] = -|\psi_1^T H_{11}|$, again by condition (ii). Also,

$$\begin{aligned}
E[\tilde{Y}_1|H_1] &= E[Y_1 + Y_2 - \mu_2(H_2, A_2)|H_1] \\
&= E[Y_2 - Q_2(H_2, A_2) + Y_1 + \max_{a_2} Q_2(H_2, a_2)|H_1] \\
&= E\Big[E[Y - Q_2(H_2, A_2)|H_2, A_2]\Big|H_1\Big] + E[Y_1 + \max_{a_2} Q_2(H_2, a_2)|H_1] \\
&= 0 + E\Big[E[Y_1 + \max_{a_2} Q_2(H_2, a_2)|H_1, A_1]\Big|H_1\Big] \\
&= E[Q_1(H_1, A_1)|H_1] \\
&= m_1(H_1).
\end{aligned}$$

Finally, plug in $Y_1 + Y_2 - \mu_2(H_2, A_2; \hat{\psi}_2)$ for $\tilde{Y}_1$. Thus, $\hat{\psi}_1$ in Robins' method solves the following reduced version of (5.20):

$$\mathbb{P}_n\left[\Big(H_{11}A_1\Big)\Big(Y_1 + Y_2 - H_{21}^T A_2 \hat{\psi}_2 + |\hat{\psi}_2^T H_{21}| - m_1(H_1) - H_{11}^T A_1 \hat{\psi}_1\Big)\right] = 0.$$

for any choice of $m_1(H_1)$ (again omitting the conditional variance). In particular, for $m_1(H_1) = H_{10}^T \hat{\beta}_1$, where $\hat{\beta}_1$ is given by (5.19), this estimating equation exactly matches with that of Q-learning.

In summary, the Q-learning algorithm as presented here is inefficient because: (a) it sets the conditional variances to be constant over (H_j, A_j), and (b) uses $H_{j1}A_j$ instead of the "efficient choice" of the term $S_{\text{eff},j}$ (that attains semiparametric variance bound) in Robins' estimating equation (see [71, p. 212]; more details in [70]).

□

5.8 Appendix B: Proof of Lemma V.2

Proof. To estimate the hyper-parameter ϕ^2, first integrate out μ to get the marginal likelihood $X|\phi^2 \sim N(0, \phi^2 + \sigma^2)$. The corresponding Jeffrey's prior on the vari-

ance parameter is $p(\phi^2) \propto 1/(\phi^2+\sigma^2)$. Based on this formulation, the posterior distribution of ϕ^2 is given by

$$p(\phi^2|X) \propto (\phi^2+\sigma^2)^{-3/2} \exp\Big\{ -\frac{X^2}{2(\phi^2+\sigma^2)}\Big\}.$$

Hence the posterior mode of ϕ^2 is

$$\widehat{\phi^2} = \arg\max_{\phi^2 \geq 0}\ p(\phi^2|X) = \left(\frac{X^2}{3} - \sigma^2\right)^+. \tag{5.21}$$

Given $\phi^2 = \widehat{\phi^2}$, now we will consider the data likelihood $X|\mu \sim N(\mu, \sigma^2)$ along with the prior $\mu|\phi^2 \sim N(0, \phi^2)$ to derive an empirical Bayes estimator for $|\mu|$. It is easy to show that the posterior distribution of μ given $\phi = \hat{\phi}$ is

$$\mu|X, \hat{\phi} \sim N\left(\frac{X\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}, \frac{\sigma^2\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}\right). \tag{5.22}$$

Now under the squared error loss, the Bayes estimator of $|\mu|$ is $E_{\mu|X}(|\mu|)$ which can be calculated using (5.22). If $Y \sim N(\theta, \tau^2)$, then $E|Y|$ is given by:

$$E|Y| = \theta\Big(2\Phi(\theta/\tau) - 1\Big) + \sqrt{\frac{2}{\pi}}\tau\ e^{-\theta^2/2\tau^2}. \tag{5.23}$$

In the present problem,

$$Y = \mu|X, \qquad \theta = \frac{X\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}, \qquad \tau^2 = \frac{\sigma^2\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}.$$

$$\text{Hence,} \qquad \frac{\theta}{\tau} = \frac{X}{\sigma}\sqrt{\frac{\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}}, \qquad \frac{\theta^2}{2\tau^2} = \frac{X^2}{2\sigma^2}\Big(\frac{\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2}\Big).$$

From (5.21), we get

$$\frac{\widehat{\phi^2}}{\widehat{\phi^2}+\sigma^2} = \frac{(X^2-3\sigma^2)^+}{X^2} = \Big(1-\frac{3\sigma^2}{X^2}\Big)^+,$$

$$\theta = X\left(1-\frac{3\sigma^2}{X^2}\right)^+, \qquad \tau^2 = \sigma^2\left(1-\frac{3\sigma^2}{X^2}\right)^+,$$

$$\frac{\theta}{\tau} = \frac{X}{\sigma}\sqrt{\Big(1-\frac{3\sigma^2}{X^2}\Big)^+}, \qquad \frac{\theta^2}{2\tau^2} = \frac{X^2}{2\sigma^2}\left(1-\frac{3\sigma^2}{X^2}\right)^+.$$

Thus an empirical Bayes estimator of $|\mu|$ is given by

$$\begin{aligned} \widehat{|\mu|}^{EB} &= X\left(1-\frac{3\sigma^2}{X^2}\right)^+\left(2\Phi\Big(\frac{X}{\sigma}\sqrt{\Big(1-\frac{3\sigma^2}{X^2}\Big)^+}\Big)-1\right) \\ &+ \sqrt{\frac{2}{\pi}}\,\sigma\sqrt{\Big(1-\frac{3\sigma^2}{X^2}\Big)^+}\exp\left\{-\frac{X^2}{2\sigma^2}\left(1-\frac{3\sigma^2}{X^2}\right)^+\right\}. \end{aligned} \tag{5.24}$$

□

5.9 Appendix C: A Very Brief Review of Multiple Imputation

Multiple imputation (MI) [75] is a simulation-based Bayesian approach to handle missing data. The basic idea is to solve an incomplete-data problem by repeatedly, say $m(>1)$ times, solving a corresponding complete-data problem, and finally combining (averaging) them. The different imputations are conceived as independent draws from a posterior predictive distribution for the missing data, given the observed data. The variation among the m imputations reflect the uncertainty in predicting the missing values from the observed ones. A single imputation cannot capture this uncertainty; this is why MI is generally considered a better missing data technique than single imputation.

Generally speaking, MI operates by assuming a model for the complete multivariate data and also assuming certain priors for the parameters of the multivariate data model. It is found that even though one assumes parametric models, the procedure is considerably robust to mis-specification of the complete data model, particularly if the "rate of missingness" is not very high. This is, at least partly, due to the fact that the mis-specification, if any, applies only to the missing part of the data, and not to the entire data (as in the EM algorithm). While analyzing data, one can use an analysis model that is quite different from the imputation model. Thus, the mis-specifications of the imputation model do not necessarily carry over to the data

analysis model. The MI procedure is often implemented via Markov Chain Monte Carlo (MCMC) techniques. However the MI procedure is still not very demanding computationally, since only a few (say, $5-10$) imputations are usually good enough. See [79] for further discussion.

To make the above discussion more precise, let us introduce some notation. Let the complete data Y be partitioned into an observed part (Y_{obs}) and a missing part (Y_{mis}), e.g., $Y = (Y_{obs}, Y_{mis})$. Let R denote the indicator of response (or equivalently, missingness). Let θ denote the parameters of the data model and ξ denote the parameters governing the missingness mechanism (e.g., the distribution of R). In this set-up, *ignorability* of the missingness mechanism means the following:

1. The data are missing at random (MAR), i.e., $P(R|Y_{obs}, Y_{mis}, \xi) = P(R|Y_{obs}, \xi)$;
2. The parameters θ and ξ are a priori independent.

We will assume ignorability throughout. As mentioned earlier, an imputation is a random draw from the posterior predictive distribution for the missing data, given the observed data:

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs}, \theta)P(\theta|Y_{obs})d\theta \tag{5.25}$$

Note that the above is an intractable integral over θ; moreover $P(\theta|Y_{obs})$ itself is an intractable integral over Y_{mis}. However it is relatively straightforward to sample from $\theta|(Y_{obs}, Y_{mis})$ and $Y_{mis}|(Y_{obs}, \theta)$. Thus to overcome the intractability of (5.25), one uses an MCMC sampling scheme to generate a draw from $Y_{mis}|Y_{obs}$. The procedure is known as *data augmentation* [92], which is a close variant of *Gibbs sampling*. The algorithm follows:

1. Start with an initial guess for θ, say $\theta^{(0)}$

2. At the $(t+1)$-th iteration, $t = 0, 1, 2, \ldots$,

$$Y_{mis}^{(t+1)} \sim Y_{mis}|Y_{obs}, \theta^{(t)},$$
$$\theta^{(t+1)} \sim \theta|Y_{obs}, Y_{mis}^{(t+1)}.$$

3. After a large number of iterations (e.g. the *burn-in* period), Y_{mis} drawn this way can be taken as a draw from the true target distribution $P(Y_{mis}|Y_{obs})$.

Note that the successive draws from the Markov chain will be correlated. To overcome this, the successive imputations should be far apart in the chain. See [79] for further details.

Imputations under a Multivariate Normal Model

Often imputations are done assuming a multivariate normal distribution for the complete data Y, e.g., $N_p(\mu, \Sigma)$ (thus $\theta = (\mu, \Sigma)$ in this case), where p is the dimension of Y. The conjugate prior for this family is normal for μ and inverted-Wishart for Σ (together called the normal inverted-Wishart prior). It is given by

$$\mu|\Sigma \sim N_p(\mu_0, \tau^{-1}\Sigma),$$
$$\Sigma \sim W^{-1}(m, \Lambda),$$

for fixed hyper-parameters $\mu_0 \in \mathbb{R}^p$, $\tau > 0$, $m \geq p$, and $\Lambda > 0$. However when no strong prior information is available about θ, it is customary to use the improper prior

$$\pi(\theta) \propto |\Sigma|^{-(\frac{p+1}{2})}, \tag{5.26}$$

which is the limiting form of the normal inverted-Wishart prior as $\tau \to \infty$, $m \to -1$ and $\Lambda^{-1} \to 0$. Note that μ does not appear on the right side of (5.26); the prior

distribution of μ is assumed to be uniform over the p-dimensional real space. Under this improper prior, the complete-data posterior becomes

$$\mu|\Sigma, Y \sim N_p(\bar{Y}, n^{-1}\Sigma),$$
$$\Sigma|Y \sim W^{-1}(n-1, (nS)^{-1}),$$

where S is the sample sums of squares and cross-products matrix. See [79] for further details.

Conditional Modeling

When some of the variables in a data set are completely observed, a better imputation strategy is to use them simply as covariates (denoted by X, say) and assume a multivariate model for the remaining variables that are subject to missingness, say Y conditionally on X. By following this strategy, one relaxes the assumption of normality for the completely observed covariates, thus lowering the risk of model mis-specification. This means that the data model becomes $y_i|x_i \sim N(\beta^T x_i, \Sigma)$, with the noninformative improper prior for β and inverted-Wishart prior for Σ. Here i denotes a certain row in the data set ($i = 1, \ldots, n$). This modeling strategy is allowed by the R package NORM Version 3, developed by Schafer [80].

Normal Modeling for Non-normal Data

Sometimes multiple imputations are conducted assuming a joint normal distribution for the data, even when there are some binary, categorical or ordinal variables in the data set [79]. In such cases, imputed values of binary or ordinal variables are rounded or classified after imputation to get a sensible imputed data set. Using a normal model for a categorical variable needs special consideration; one reasonable

strategy is to first convert the categorical variable into several binary variables, use a normal approximation for each of them, and then round off the value to make it binary (after imputation). Robustness of such model mis-specifications was studied by Bernaards et al. [4].

Our smoking cessation data set contains binary, categorical and ordinal variables along with continuous variables. None-the-less we used the multivariate normal modeling because of its conceptual simplicity, ready availability of software and its allowance of conditional modeling. We used some ad hoc but sensible strategies [79, 80, 4] to make post-imputation adjustments (rounding, classification etc.).

CHAPTER VI

Future Work and Conclusion

This dissertation explores two research projects, e.g. the problem of designing multicomponent intervention trials, and the problem of non-regularity in the dynamic treatment regime framework. Each of these projects investigated in the previous chapters has its own direction; each raise new and challenging questions and opens up possibilities for further research. In the following we briefly discuss some of these problems that we plan to explore in greater depth in future.

6.1 Follow up Studies for Multicomponent Interventions

In Chapter III, we discussed the usefulness of FFDs in screening trials for developing multicomponent interventions. However, screening study is only the first phase of the MOST framework. Depending on the aliasing pattern of the screening design used, one often needs to conduct follow-up or refining studies to resolve any remaining research questions regarding the various components and their interactions (e.g. to de-alias significant aliased interactions) after the completion of the screening study. In Chapters III and IV, we provided some examples of follow-up studies. However some principled approach needs to be developed in this area. For example, some Bayesian approach, where one puts suitable priors on the effects of various

components and interactions, might be useful for designing follow-up studies. Such an approach can be found in [52]. Further strategies for conducting follow-up studies can be found in [107]. Also, in case there is at least one component with more than two levels (e.g. a continuous component), dose-response experiments where subjects are randomized to ethical doses should be used to find the optimal dose of these components. The existing literature on *response surface experiments* [10, 61] may prove useful for finding designs applicable to the multicomponent intervention setting. We would like to explore these areas to develop efficient follow-up study designs in future.

6.2 Soft-threshold Estimator for More than Two Treatments per Stage

In Chapter V, we discussed the soft-threshold estimator in the context of Q-learning. However, we assumed only two treatment options per stage all along. A natural question that arises here is how to generalize this estimator in the setting of more than two treatments per stage. We do not have a direct answer at this point.

Recall that the problem of non-regularity occurs due to a non-smooth maximum operation, e.g. $\max_a Q_2(H_2, a)$. When the Q-functions are linear (as in this dissertation), maximization over a binary a gives rise to a piecewise linear function with only one hinge or point of non-differentiability (see Chapter V for details). When there are more than two treatments per stage, i.e., when a takes more than two values, $\max_a Q_2(H_2, a)$ will be a piecewise linear function with possibly more than one point of non-differentiability.

Now in case of two treatments per stage, the soft-threshold estimator works by thresholding this piecewise linear function around its point of non-differentiability. To extend this idea to the multi-treatment case, one would like to somehow threshold

the function $\max_a Q_2(H_2, a)$ around each of its points of non-differentiability. Note, however, that it is not even clear what thresholding means in such a scenario. Even when thresholding has some meaning, it is not clear how to choose the multiple tuning parameters, one for each point of non-differentiability. If one takes recourse to a Bayesian approach, some suitable multivariate prior needs to chosen. This, clearly, is a non-trivial problem; some serious thought is needed. We would like to investigate this in future.

6.3 Consistent Bootstrap Procedure for Non-regular Settings

It is well-known [84, 2] that the usual bootstrap procedure is inconsistent for non-regular (non-differentiable) settings. Note that even though the threshold estimators (both hard and soft) empirically seemed to perform better than the original hard-max estimator (see Chapter V), both the hard-threshold and the soft-threshold estimators are non-regular (as is the original hard-max estimator). Hence the validity of usual bootstrap CIs for threshold estimators is not theoretically justified. A nice theoretical project would be to develop a consistent bootstrap procedure, e.g. some "adaptive" version of the usual bootstrap, to use in conjunction with the soft-threshold estimator. We would like to explore this in future.

6.4 Concluding Remarks

This dissertation investigated two broad problems with practical relevance to clinical trials and medical statistics. "Evidence-based" treatments (treatment sequences) are of great interest in the behavioral sciences and in medicine; our research is an ongoing endeavor to address that interest. In particular, the area of dynamic treat-

ment regimes is very young and exciting; and there are many interesting problems in this area, some of which we want to address in near future.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] H.G. Allore, M.E. Tinettia, T.M. Gill, and P.N. Peduzzi. Experimental designs for multicomponent interventions among persons with multifactorial geriatric syndromes. *Clinical Trials*, 2(1):13–21, 2005.

[2] D.W.K. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399 – 405, 2000.

[3] B.E. Ankerman, A.I. Aviles, and J.C. Pinheiro. Optimal designs for mixed-effects models with two random nested factors. *Statistica Sinica*, 13:385–401, 2003.

[4] C.A. Bernaards, T.R. Belin, and J.L. Schafer. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26:1368–1382, 2007.

[5] P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, 1993.

[6] P. Bickel and A. Sakov. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, 18(3):967 – 985, 2008.

[7] D. Blatt, S.A. Murphy, and J. Zhu. A-learning for approximate planning. *Technical Report 04-63, The Methodology Center, Pennsylvania State University*, 2004.

[8] D.A. Bluford, B. Sherry, and K.S. Scanlon. Interventions to prevent or treat obesity in preschool children: A review of evaluated programs. *Obesity*, 15:1356–72, 2007.

[9] G. Box, W. Hunter, and J. Hunter. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model building*. Wiley, New York, 1978.

[10] G.E.P. Box and N.R. Draper. *Empirical Model-building and Response Surfaces*. Wiley, New York, 1987.

[11] G.E.P. Box and J.S. Hunter. The 2^{k-p} fractional factorial designs. *Technometrics*, 3:311–351 and 449–458, 1961.

[12] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373 – 384, 1995.

[13] E. Brittain and J. Wittes. Factorial designs in clinical trials: the effects of non-compliance and subadditivity. *Statistics in Medicine*, 8:161–171, 1989.

[14] D.P. Byar. Factorial and reciprocal control designs (with discussion). *Statistics in Medicine*, 9:55–64, 1990.

[15] D.P. Byar, A.M. Herzberg, and W. Tan. Incomplete factorial designs for randomized clinical trials. *Statistics in Medicine*, 12:1629 – 1641, 1993.

[16] D.P. Byar and S. Piantadosi. Factorial designs for randomized clinical trials. *Cancer Treatment Reports*, 69:1055–1063, 1985.

[17] K. Byth and V. Gebski. Factorial designs: a graphical aid for choosing study designs accounting for interaction. *Clinical Trials*, 1:315325, 2004.

[18] M. Campbell, R. Fitzpatrick, A. Haines, A.L. Kinmonth, P. Sandercock, D. Spiegelhalter, and P. Tyrer. Framework for design and evaluation of complex interventions to improve health. *British Medical Journal*, 321:694 – 696, 2000.

[19] B. Chakraborty, L. Collins, V. Strecher, and S. Murphy. Developing multicomponent interventions using fractional factorial designs. *To appear in Statistics in Medicine*, 2009.

[20] B. Chakraborty, V. Strecher, and S. Murphy. Inference for non-regular parameters in optimal dynamic treatment regimes. *To appear in Statistical Methods in Medical Research*, 2009.

[21] J. Cohen. *Statistical Power for the Behavioral Sciences*. Erlbaum, Hillsdale, NJ, 2nd edition, 1988.

[22] L.M. Collins, B. Chakraborty, S.A. Murphy, and V.J. Strecher. Comparison of a phased experimental approach and a single randomized clinical trial for developing multicomponent behavioral interventions. *Clinical Trials*, 6(1):5–15, 2009.

[23] L.M. Collins, J.J. Dziak, and R. Li. Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *To appear in Psychological Methods*, 2009.

[24] L.M. Collins, S.A. Murphy, and K. Bierman. A conceptual framework for adaptive preventive interventions. *Prevention Science*, 5:185 – 196, 2004.

[25] L.M. Collins, S.A. Murphy, V.N. Nair, and V.J. Strecher. A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine*, 30:65–73, 2005.

[26] D.J. Couper, J.D. Hosking, R.A. Cisler, D.R. Gastfriend, and D.R. Kivlahan. Factorial designs in clinical trials: Options for combination treatment studies. *Journal of Studies on Alcohol*, S15:24–32, 2005.

[27] CPPRG. A developmental and clinical model for the prevention of conduct disorders: The fast track program. *Development and Psychopathology*, 4:509–528, 1992.

[28] M. Cuffe. The patient with cardiovascular disease: treatment strategies for preventing major events. *Clinical Cardiology*, 29:II4–12, 2006.

[29] R. Curhan. The effects of merchandising and temporary promotional activities on the sales of fresh fruits and vegetables in supermarkets. *Journal of Marketing Research*, 11(3):286–294, 1974.

[30] A.P. Daunica, S.W. Smitha, E.M. Branka, and R.D. Penfield. Classroom-based cognitivebehavioral intervention to prevent aggression: Efficacy and social validity. *Journal of School Psychology*, 44(2):123–139, 2006.

[31] A.C. Davison and D.V. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, Cambridge, UK, 1997.

[32] R. Dawson and P.W. Lavori. Placebo-free designs for evaluating new mental health treatments: the use of adaptive treatment strategies. *Statistics in Medicine*, 23:3249 – 3262, 2004.

[33] A. Dijkstra, H. DeVries, J. Roijackers, and G. Van Breukelen. Tailored interventions to communicate stage-matched information to smokers in different motivational stages. *Journal of Consulting and Clinical Psychology*, 66(3):549–557, 1998.

[34] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425 – 455, 1994.

[35] M. Figueiredo and R. Nowak. Wavelet-based image estimation: an empirical bayes approach using jeffreys' noninformative prior. *IEEE Transactions on Image Processing*, 10(9):1322 – 1331, 2001.

[36] R. Fisher. *The Design of Experiments, 3rd ed.* Oliver & Boyd, Edinburgh, 1942.

[37] B.R. Flay and L.M. Collins. Historical review of school-based randomized trials for evaluating problem behavior prevention programs. *Annals of the American Academy of Political and Social Science*, 599:115–146, 2005.

[38] B. Freedman. Equipoise and the ethics of clinical research. *The New England Journal of Medicine*, 317(3):141 – 145, 1987.

[39] K. Friedli and M.B. King. Psychological treatments and their evaluation. *International Review of Psychiatry*, 10:123 – 126, 1998.

[40] H. Gao. Wavelet shrinkage denoising using the nonnegative garrote. *Journal of Computational and Graphical Statistics*, 7:469 – 488, 1998.

[41] C.E. Golin, J. Earp, H.C. Tien, P. Stewart, C. Porter, and L. Howie. A 2-arm, randomized, controlled trial of a motivational interviewing-based intervention to improve adherence to antiretroviral therapy (art) among patients failing or initiating art. *Journal of Acquired Immune Deficiency Syndrome*, 42:42–51, 2006.

[42] S. Green, P.Y. Liu, and J. O'Sullivan. Factorial design considerations. *Journal of Clinical Oncology*, 20(16):3424 – 3430, 2002.

[43] P. Hall, J. Horowitz, and B. Jing. On blocking rules for the bootstrap with dependent data. *Biometrika*, 82:561 – 574, 1995.

[44] P. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–970, 1986.

[45] J.D. Hosking, R.A. Cisler, D.J. Couper, D.R. Gastfriend, D.R. Kivlahan, and R.F. Anton. Design and analysis of trials of combination therapies. *Journal of Studies on Alcohol*, S15:34–42, 2005.

[46] I.M. Johnstone and B.W. Silverman. Needles and straw in haystackes: empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594 – 1649, 2004.

[47] A.E. Kazdin. The evaluation of psychotherapy: Research design and methodology. In S.L. Garfield and A.E. Bergin, editors, *Handbook of psychotherapy and behavior change*, pages 23 – 68, New York, 1986. Wiley.

[48] E.L. Korn, D.M. Teeter, and S. Baumrind. Using explicit clinician preferences in nonrandomized study designs. *Journal of Statistical Planning and Inference*, 96:6782, 2001.

[49] P.W. Lavori. A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society, Series A*, 163:29 – 38, 2000.

[50] P.W. Lavori and R. Dawson. Dynamic treatment regimes: practical design considerations. *Clinical Trials*, 1:9 – 20, 2004.

[51] J.K. Lunceford, M. Davidian, and A.A. Tsiatis. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 58:48 – 57, 2002.

[52] R.D. Meyer, D.M. Steinberg, and G.E.P. Box. Follow-up designs to resolve confounding in multifactor experiments. *Technometrics*, 38:303–332, 1996.

[53] A.A. Montgomery, T.J. Peters, and P. Little. Design, analysis and presentation of factorial randomized controlled trials. *BMC Medical Research Methodology*, 3(26), 2003.

[54] E.E.M. Moodie and T.S. Richardson. Estimating optimal dynamic regimes: correcting bias under the null. *To appear in Scandinavian Journal of Statistics*, 2008.

[55] E.E.M. Moodie, T.S. Richardson, and D.A. Stephens. Demystifying optimal dynamic treatment regimes. *Biometrics*, 63:447–455, 2007.

[56] S.A. Murphy. Optimal dynamic treatment regimes (with discussions). *Journal of the Royal Statistical Society, Series B*, 65:331–366, 2003.

[57] S.A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24:1455–1481, 2005.

[58] S.A. Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6:1073–1097, 2005.

[59] S.A. Murphy and D. Bingham. Screening experiments for developing adaptive treatment strategies. *To appear in the Journal of the American Statistical Association*, 2008.

[60] S.A. Murphy, M.J. van der Laan, J.M. Robins, and CPPRG. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96:1410–1423, 2001.

[61] R.H. Myers and D.C. Montgomery. *Response Surface Methodology.* Wiley, New York, 1995.

[62] V. Nair, V. Strecher, A. Fagerlin, P. Ubel, K. Resnicow, S. Murphy, R. Little, B. Chakraborty, and A. Zhang. Screening experiments and fractional factorial designs in behavioral intervention research. *American Journal of Public Health*, 98:1354 – 1359, 2008.

[63] J.C. Nankervis. Computational algorithms for double bootstrap confidence intervals. *Computational Statistics & Data Analysis*, 49:461–475, 2005.

[64] G. Parmigiani. *Modeling in medical decision making: a Bayesian approach.* John Wiley & Sons, New York, 2002.

[65] G. Paul, S.M. Smith, D. Whitford, F. O'Kelly, and T. O'Dowd. Development of a complex intervention to test the effectiveness of peer support in type 2 diabetes. *BMC Health Services Research*, 7:136, 2007.

[66] J. Pearl. Graphs, causality, and structural equation models. *Sociological Methods and Research*, 27(2):226–84, 1998.

[67] S. Piantadosi. *Clinical Trials: A Methodologic Perspective.* Wiley, New York, 2005.

[68] N.R. Riggs, P. Elfenbaum, and M.A. Pentz. Parent program component analysis in a drug abuse prevention trial. *Journal of Adolescent Health*, 39:66 – 72, 2006.

[69] D.E. Rivera, M.D. Pew, and L.M. Collins. Using engineering control principles to inform the design of adaptive interventions: a conceptual introduction. *Drug and Alcohol Dependence*, 88:S31S40, 2007.

[70] J.M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics*, 23:2379–2412, 1994.

[71] J.M. Robins. Optimal structural nested models for optimal sequential decisions. In D.Y. Lin and P. Heagerty, editors, *Proceedings of the Second Seattle Symposium on Biostatistics*, pages 189–326, New York, 2004. Springer.

[72] P.R. Rosenbaum and D.B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524, 1984.

[73] S. Rosthoj, C. Fullwood, R. Henderson, and S. Stewart. Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach. *Statistics in Medicine*, 25:4197–4215, 2006.

[74] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[75] D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

[76] V. Rukshin, R. Santos, and et al. A prospective, non-randomized, open-labeled pilot study investigating the use of magnesium in patients undergoing nonacute percutaneous coronary intervention with stent implantation. *Journal of Cardiovascular Pharmacology and Therapeutics*, 8:193200, 2003.

[77] A.J. Rush, M.L. Crismon, and et al. Texas medication algorithm project, phase 3 (tmap-3): Rationale and study design. *Journal of Clinical Psychiatry*, 64(4):357–369, 2003.

[78] A.J. Rush, M. Fava, S.R. Wisniewski, P.W. Lavori, M.H. Trivedi, and H.A. Sackeim. Sequenced treatment alternatives to relieve depression (star*d): rationale and design. *Controlled Clinical Trials*, 25:119–142, 2003.

[79] J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.

[80] J.L. Schafer. *NORM: Analysis of incomplete multivariate data under a normal model, Version 3. Software Package for R*. The Methodology Center, The Pennsylvania State University, University Park, PA, 2008.

[81] L.S. Schneider, P.N. Tariot, C.G. Lyketsos, K.S. Dagerman, K.L. Davis, and S. Davis. National institute of mental health clinical antipsychotic trials of intervention effectiveness (catie): alzheimer disease trial methodology. *American Journal of Geriatric Psychology*, 9:346–360, 2001.

[82] S.R. Searle, G. Casella, and C.E. McCulloch. *Variance Components*. Wiley, New York, 2002.

[83] W.R. Shadish, T.D. Cook, and D.T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin, New York, 2002.

[84] J. Shao. Bootstrap sample size in nonregular cases. *Proceedings of the American Mathematical Society*, 122(4):1251–1262, 1994.

[85] J.R. Smith and J.M. Beverly. The use and analysis of staggered nested factorial designs. *Journal of Quality Technology*, 13:166–173, 1981.

[86] J. Stephenson and J. Imrie. Why do we need randomised controlled trials to assess behavioural interventions? *British Medical Journal*, 316:611 – 613, 1998.

[87] V. Strecher. Computer-tailored smoking cessation materials: a review and discussion. *Patient Education and Counselling*, 36:107–117, 1999.

[88] V. Strecher, J. McClure, G. Alexander, B. Chakraborty, V. Nair, J. Konkel, S. Greene, L. Collins, C. Carlier, C. Wiese, R. Little, C. Pomerleau, and O. Pomerleau. Web-based smoking cessation components and tailoring depth: Results of a randomized trial. *American Journal of Preventive Medicine*, 34(5):373 – 381, 2008.

[89] V. Strecher, S. Shiffman, and R. West. Randomized controlled trial of a web-based computer-tailored smoking cessation program as a supplement to nicotine patch therapy. *Addiction*, 100:682–688, 2005.

[90] R.S. Sutton and A.G. Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, 1998.

[91] L. Swartz, J. Noell, S. Schroeder, and D. Ary. A randomized control study of a fully automated internet based smoking cessation programme. *Tobacco Control*, 15:7–12, 2006.

[92] M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528–550, 1987.

[93] P.F. Thall, R.E. Millikan, and H.G. Sung. Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine*, 30:1011–1128, 2000.

[94] P.F. Thall, H.G. Sung, and E.H. Estey. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association*, 97(457):29–39, 2002.

[95] P.F. Thall, L.H. Wooten, C.J. Logothetis, R.E. Millikan, and N.M. Tannir. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in Medicine*, 26:4687–4702, 2007.

[96] M.E. Tinetti, D.I. Baker, G. McAvary, and et al. A multifactorial intervention to reduce the risk of falling among elderly people living in the community. *New England Journal of Medicine*, 331:821–827, 1994.

[97] M.E. Tinetti, G. McAvary, and E. Claus. Does multiple risk factor reduction explain the reduction in fall rate in the yale ficsit trial? *American Journal of Epidemiology*, 144:389–399, 1996.

[98] W.P. Vogt. *Dictionary of statistics and methodology: A nontechnical guide for the social sciences*. SAGE, Newbury Park, CA, 1993.

[99] A.S. Wahed and A.A. Tsiatis. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomized designs in clinical trials. *Biometrics*, 60:124 – 133, 2004.

[100] A.S. Wahed and A.A. Tsiatis. Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. *Biometrika*, 93(1):163 – 177, 2006.

[101] C.J.C.H. Watkins. Learning from delayed rewards. *Ph.D. Thesis, Cambridge University*, 1989.

[102] S.G. West and L.S. Aiken. Toward understanding individual effects in multicomponent prevention programs: Design and analysis strategies. In K.J. Bryant, M. Windle, and S.G. West, editors, *The Science of Prevention: Methodological Advances From Alcohol and Substance Use Research*, Washington, DC, 1997. American Psychological Association.

[103] S.G. West, L.S. Aiken, and M. Todd. Probing the effects of individual components in multiple component prevention programs. *American Journal of Community Psychology*, 21(5):571–605, 1993.

[104] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817838, 1980.

[105] J.W. Williams, M. Gerrity, T. Holsinger, S. Dobscha, B. Gaynes, and A. Dietrich. Systematic review of multifaceted interventions to improve depression care. *General Hospital Psychiatry*, 29:91–116, 2007.

[106] S.A. Wolchik, S.G. West, S. Westover, I.N. Sandler, A. Martin, J. Lustig, J.Y. Tein, and J. Fisher. The children of divorce intervention project: Outcome evaluation of an empirically based parenting program. *American Journal of Community Psychology*, 21:293–331, 1993.

[107] C.F.J. Wu and M. Hamada. *Experiments: Planning, Analysis, and Parameter Design Optimization*. Wiley, New York, 2000.

[108] F. Yates. The design and analysis of factorial experiments. *Imperial Bureau of Soil Sciences - Technical communication No. 35*, 1937.

[109] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418 – 1429, 2006.