Summer 8-6-2019

Clinical Trial Design and Analysis

Shuang Li
Southern Methodist University, shuangli@smu.edu

CLINICAL TRIAL DESIGN AND ANALYSIS

Approved by:

_____

Dr. Daniel F. Heitjan
Professor in Department of Statistical Sciences, SMU, and Population & Data Sciences, UTSW

_____

Dr. Xinlei (Sherry) Wang
Professor in Department of Statistical Sciences, SMU

_____

Dr. Xian-Jin Xie
Professor in Division of Biostatistics and Computational Biology, College of Dentistry, and Department of Biostatistics, College of Public Health, University of Iowa

_____

Dr. Jarett Berry
Associate Professor in Department of Internal Medicine, UTSW

_____

Dr. Song Zhang
Associate Professor in Department of Population & Data Sciences, UTSW

CLINICAL TRIAL DESIGN AND ANALYSIS

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Statistical Science

by

Shuang Li

B.S., Mathematics, University of Utah

August 6, 2019

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Dr. Daniel F. Heitjan, for his mentoring during the past two years when I worked on my dissertation. Dr. Heitjan has set a great example showing me how to be a statistician, and how to present the materials that I have been working on for a long time to someone who has little background knowledge. With his help, I have received an award for Biopharm Scholar at the 74th Annual Deming Conference in December 2018.

I also appreciate working with Dr. Xian-Jin Xie on part of my dissertation about dose-finding designs for phase I oncology trials. He has provided many helpful ideas about the area. And I am also grateful for the rest of my committee members: Dr. Xinlei (Sherry) Wang, Dr. Jarett Berry, and Dr. Song Zhang. They have given thoughtful comments and suggestions to help improve my dissertation.

In addition, I would like express my gratitude to my parents and family members living in Salt Lake City, Utah. I would not have received this PhD degree without their support.

Last but not at least, many thanks to my boyfriend Dr. Yu Lan for his love and encouragement to share with me my ups and downs for the past three years.

Li, Shuang B.S., Mathematics, University of Utah

Clinical Trial Design and Analysis

Advisor: Dr. Daniel F. Heitjan

Doctor of Philosophy degree conferred August 6, 2019
Dissertation completed July 22, 2019

Clinical trials are experiments tested on human to compare the effect of certain intervention. Any intervention has to be tested in different phases of clinical trials. In early-stage (phase 0, I, and II) trials, fewer number of patients are enrolled to get preliminary information on safety and efficacy. In late-stage (phase III and IV) trials, larger number of patients are randomized to further confirm the efficacy and safety. The purpose of each phase may change depending on the therapeutic area and patient characteristics.

In Chapter 2, we propose a family of designs for phase I oncology trials. In these trials, oncologists assign different patients at a varying range of dose levels to find the dose that gives the highest acceptable rate of dose-limiting toxicities, which will be the recommended dose for phase II trials. In current practice, most such trials are rule-based designs that determine whether to escalate the dose using data from the current dose only. The 3+3 design is the most popular rule-based design. Our proposed design, which we denote the cohort-sequence design, addresses the deficiencies of the 3+3 design, while preserving its simplicity. We compare the proposed design with the 3+3 using simulation studies.

Late-stage randomized clinical trials might lack external validity when there are treatment-by-covariate interactions involving factors whose distribution in the population differ from that in the trial. In Chapter 3, we project the results from a trial to a population using post-stratification, and compare it with other approaches that use probabilities of trial inclusion. We use intention-to-treat estimate of the treatment effect, which compares the difference in average responses in assigned experimental treatment versus control groups. We demonstrate

this using data from Lipids Research Clinics Coronary Primary Prevention Trial.

The intention-to-treat analysis focuses on the treatment effect of randomization on the outcome, without considering compliance. In Chapter 4, we extend the interpolation approaches using instrumental variables estimation of the complier average causal effect, which is the treatment effect on the outcome restricted to the subset of subjects who adhere to assigned treatment. We apply these methods to the data from Lipids Research Clinics Coronary Primary Prevention Trial and New York School Choice Experiment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I dedicate this dissertation to my parents, Xiaoping Wang and Baokui Li; my sister Wen Li; and my family in Utah, Chaoying Li, Yong Zeng, and Jiangmei Wang.

CHAPTER 1

INTRODUCTION

1.1. Background

Clinical trials are experiments tested on human beings to compare the effect and safety of intervention against a control. Any new treatment needs to pass different phases of clinical trials before being allowed to go into market. Phase 0 studies are proof-of-principle, studying pharmacokinetic and pharmacodynamic of an intervention. Oncology phase I trials are for dose-finding and getting preliminary information on safety. Phase II trials obtain further information on efficacy and safety. Phase III trials involve a large number of patients randomized to compare the experimental intervention with control, or standard-of-care. Phase IV trials monitor safety efficacy for the post-marking product. The purpose of each phase may vary based on the therapeutic area and diseased population.

In this dissertation, one chapter is focused on phase I dose-finding designs. The other two chapters are aiming to project the results from a randomized clinical trial to a target diseased population.

1.2. Flexible, rule-based dose escalation: The cohort-sequence design

Phase I oncology trials obtain preliminary information on the safety of novel treatments. Most trials employ rule-based designs with pre-specified escalation/de-escalation rules based on observed number of toxicities from the current dose. The objective is to identify the

maximum tolerated dose (MTD)— the highest dose with rate of dose-limiting toxicities below a pre-specified threshold. The 3+3 design is the most common rule-based design largely due to its simplicity, (28) but it has several deficiencies. First, the design is inflexible for different toxicity threshold. (22) Second, the design is not efficient, with most patients assigned at low doses with little therapeutic benefits. Third, the number of patients treated at the maximum tolerated dose is 6, so it will obtain limited information.

In Chapter 2, we devise a family of rule-based dose-finding design that avoids the problem of the 3+3, while preserving its simplicity. Our design, which we called the cohort-sequence design, involves a sequence of cohort sizes and corresponding critical values for dose-limiting toxicities. It improves the efficiency by enrolling more patients around doses that are likely to be the maximum tolerated dose. We use simulation studies to compare the cohort-sequence design with the 3+3 design.

1.3. Projecting clinical trial results to a target population

In randomized trials, the results lack external validity if there are treatment-by-covariate interactions involving factors that are represented differently from the target population. If information on the covariates in the target population is available, and the treatment effect is similar across subgroups in the trial and target population, then one can interpolate the treatment effect to the target population.

One interpolation approach is post-stratification, or re-weighting of subgroup estimates based on the covariate distribution in the target population. This approach does not require individual level data for the target population, and works well with few number of confounders. We use dataset from Lipids Research Clinics Coronary Primary Prevention Trial to demonstrate post-stratification, (30) and compare it with two other approaches using trial inclusion probabilities. (38, 48)

1.4. Projecting instrumental variable results to a target population

The intention-to-treat analysis is commonly used for measuring the treatment effect. This as-randomized analysis does not consider compliance. In Chapter 4, we extend these interpolation analyses to instrumental variables estimation of the complier average causal effect (CACE), which is the treatment effect on the outcome among patients who adhere to the treatment.

We use the Lipids Research Clinics Coronary Primary Prevention Trial, measuring compliance as the fraction of assigned packets taken. (30) In addition, we use data from New York School Choice Experiment, which is a randomized encouragement to study the effect of private school voucher on improvement in test scores. (5, 18) We use both randomization assignment and actual attendance of private school for estimating the CACE.

CHAPTER 2

FLEXIBLE, RULE-BASED DOSE ESCALATION: THE COHORT-SEQUENCE DESIGN

2.1. Introduction

Oncologists acquire preliminary information on the safety of a novel treatment through the conduct of a *phase I trial*. Whereas statisticians commonly formulate the objective of such a trial as the identification of the treatment's maximum tolerated dose (MTD) — that is, the dose that gives the highest acceptable rate of dose-limiting toxicities (DLTs). The typical design calls for enrolling subjects in dose cohorts, starting from a low dose that is believed to be safe and increasing after each cohort until encountering a designated level of toxicity.

Most phase I trials employ *rule-based* designs that use data from only the current cohort to decide what the next step will be. The most popular such design is the 3+3, a variant of the up-and-down rule.(6, 13, 24, 25, 28, 54) Some alternatives to the 3+3 include the A+B design,(54) which generalizes the cohort sizes in the 3+3; the accelerated titration design, which starts with one-patient cohorts and reverts to the 3+3 plan once toxicities appear;(46) and the toxicity probability interval (TPI) and modified TPI (mTPI) designs, which use Bayesian criteria to select the next dose.(23, 24, 25) Yet despite the availability of these and other alternative designs, many with excellent statistical properties, as recently as 2009 nearly 97% of phase I trials used the 3+3.(28)

The 3+3 design is simple and familiar but has notable deficiencies: First, for commonly encountered toxicity profiles, the 3+3 most often selects doses with DLT rates in the range

20%–25%, well below the typical nominal target of 30%–35%.(22) Second, most patients in a 3+3 trial receive doses that have low rates of both toxicity and therapeutic effect; a more efficient design would escalate quickly past these to reach doses that are of greater interest. Third, the maximum number of patients that a 3+3 enrolls at the final dose is 6, implying that it will obtain only limited information about toxicity and efficacy at the purported MTD. Thus it has become common to augment the phase I trial with a *dose-expansion cohort* — an additional group of patients who receive treatment at the identified MTD. Often, the choice of sample size for this additional cohorts is arbitrary,(32) and in any event the benefits of an expansion cohort are limited if the trial can potentially mis-estimate the MTD.

An alternative is the *model-based* design, in which one assumes an underlying parametric dose-response model and uses the accumulated data to estimate parameters and, thereby, the MTD.(4, 6, 9, 19, 28, 39, 41, 51) Some pharmaceutical companies now use such designs routinely, often in "bucket" trials or phase I/II designs. The large majority of phase I cancer trials, however, continue to employ rule-based designs, primarily the 3+3.(31) This reluctance to adopt the newer methods may reflect several factors: The substantial cost and complexity of implementing the model-based methods; lack of familiarity with them; or simply a conviction that "better is the enemy of good enough". (34)

The objective in this chapter is to devise a rule-based dose-finding design that avoids the problems of the 3+3 while preserving, to the extent possible, its simplicity. The method, which we denote the *cohort-sequence* design, permits tuning of the cohort sizes and critical values to reflect the targeted DLT rate. It improves efficiency by focusing enrollment at doses where toxicities are likely to occur, thereby creating larger cohorts in the vicinity of the MTD and obviating the need to append an arbitrarily sized dose-expansion cohort.

2.2. Methods

2.2.1. Escalation plan for the cohort-sequence design

The *cohort-sequence* design consists of a sequence of cohort sizes $n = (n_1, \ldots, n_J)$ and corresponding DLT critical values $b = (b_1, \ldots, b_J)$ that indicate whether to escalate, de-escalate, or add more subjects at the current dose. The notion is to begin with a small cohort size n_1 and escalate through the planned series of increasing doses $D = (D_1, \ldots, D_m)$, raising the cohort size as we begin to encounter toxicities. Specifically, when enrolling subjects at dose D_i with cohort size n_j, the decision to escalate, add more at the current dose, or de-escalate hinges on whether the observed number of DLTs in the cohort falls below, equals, or exceeds the corresponding critical value b_j. If the number of DLTs at dose D_i exceeds b_j, we stop treating the cohort at that dose to avoid excessive toxicities. A possible value for the sequence of cohort sizes would be $n = (1, 3, 5, 8, 10)$, with corresponding sequence of critical values $b = (1, 2, 3, 4, 5)$.

Figure 2.1 displays the flow chart for our design. Suppose that the current dose level is D_i and that our current cohort size is n_j with corresponding critical value b_j. We enroll up to n_j subjects at this dose and observe the number of DLTs as X_i. If $X_i < b_j$, we deem the current dose to be safe, and we escalate and enroll the next cohort at dose D_{i+1} with the same cohort size n_j and critical value b_j. If $X_i > b_j$, we deem the current dose unsafe, and we enroll the next cohort at the next lower dose D_{i-1} with the terminal cohort size n_J and corresponding critical value b_J. If $X_i = b_j$, we deem the current dose as potentially, but not certainly, toxic, and we increase the cohort size to n_{j+1} and the corresponding critical value to b_{j+1}, enrolling additional subjects at this dose until the cohort is filled and we can again evaluate the safety data. The escalation/de-escalation continues in this way until a safe dose is achieved ($X_i < b_J$) after a de-escalation, or the trial de-escalates to the lowest dose. If we

escalate to the highest dose D_m with cohort size n_j, and observe $X_m < b_j$, then we increase the cohort size to n_J and the critical value to b_J. We estimate the MTD as the highest safe dose evaluated. Alternatively, one can specify a total number of patients and stop when all have received treatment, again identifying the highest safe dose as the MTD.

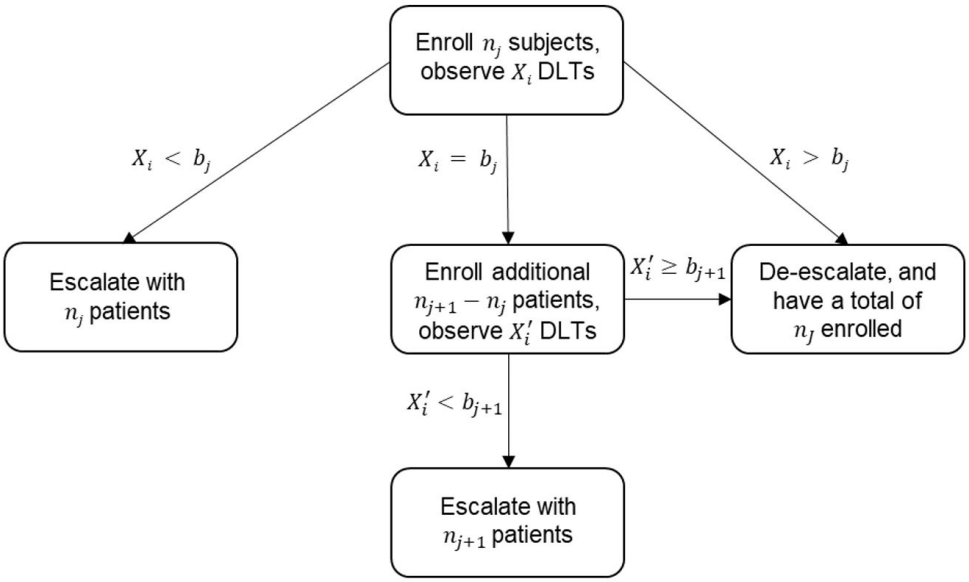


Figure 2.1. Flow chart for the cohort-sequence design.

2.2.2. Identification of the critical value for a designated cohort size

We select the pairs $(n_j, b_j),\ j = 1, \ldots, J$ to represent a comparable level of certainty about the DLT rate at the given dose. That is, for each n_j, the selected b_j is one that identifies the current dose as safe, unsafe, or indeterminate, by a criterion that is common across cohort sizes. To accomplish this, we first select a safety threshold, denoted $\theta \in (0, 1)$. The principle is that if we are reasonably sure that the DLT rate at dose D_i is below θ, we escalate; if we are reasonably sure that the DLT rate at D_i exceeds θ, we de-escalate; otherwise, we collect more data at D_i. Having chosen θ, we can either compute the critical value b_j from a selected cohort size n_j, or *vice versa*.

We first demonstrate the computation of b_j from n_j. We evaluate uncertainty about the toxicity level using Bayesian posterior probabilities. At dose level D_i, the number of DLTs X_i is binomial with parameters (N_i, τ_i), where N_i and $\tau_i = \tau(D_i)$ represent the cohort size and toxicity probability, respectively. We assume a Beta(1,4) prior for τ_i, which updates to Beta$(1 + X_i, 4 + N_i - X_i)$ after X_i DLTs in N_i trials. We use this prior because it represents the situation where there are $X = 0$ events in a previous $N = 3$ patients starting from a uniform prior, which approximates the level of uncertainty that one would express before examining an untested dose level. As in the TPI and mTPI designs,(23, 24, 25) we use the same prior for all dose levels because we have essentially no prior data at any dose level.

Starting from a fixed n_j, to determine the critical level b_j we compute the posterior probability that τ_i exceeds the safety threshold θ, which we denote

$$f(X_i, N_i, \theta) \equiv \Pr[\tau_i > \theta | X_i, N_i].$$

We fix a threshold of 10% for this posterior probability. For example, if $f(X_i, N_i, \theta)$ is below 10%, then we deem the dose level safe for escalation. If $f(X_i, N_i, \theta)$ is well above 10%, we de-escalate. If $f(X_i, N_i, \theta)$ is above 10% but $f(X_i - 1, N_i, \theta)$ is below 10%, we collect more data at the current dose. We select b_j by identifying the value of X_i such that $f(X_i, N_i, \theta) > 10\%$ and $f(X_i - 1, N_i, \theta) \leq 10\%$.

2.2.3. Identification of the cohort size for a designated critical value

Alternatively, one can select a sequence of b_j values and then calculate the corresponding cohort sizes n_j. To avoid ambiguities, the b_j values should constitute an increasing sequence; $b = (1, 2, \ldots, J)$ is a natural choice. We start by computing $f(b_j - 1, x, \theta)$ for $x \geq b_j$, which decreases as x increases. Then n_j is the smallest x such that $f(b_j - 1, x, \theta) \leq 10\%$, and $f(b_j, x, \theta) > 10\%$. Figure 2.2 illustrates the process of identifying the cohort size.

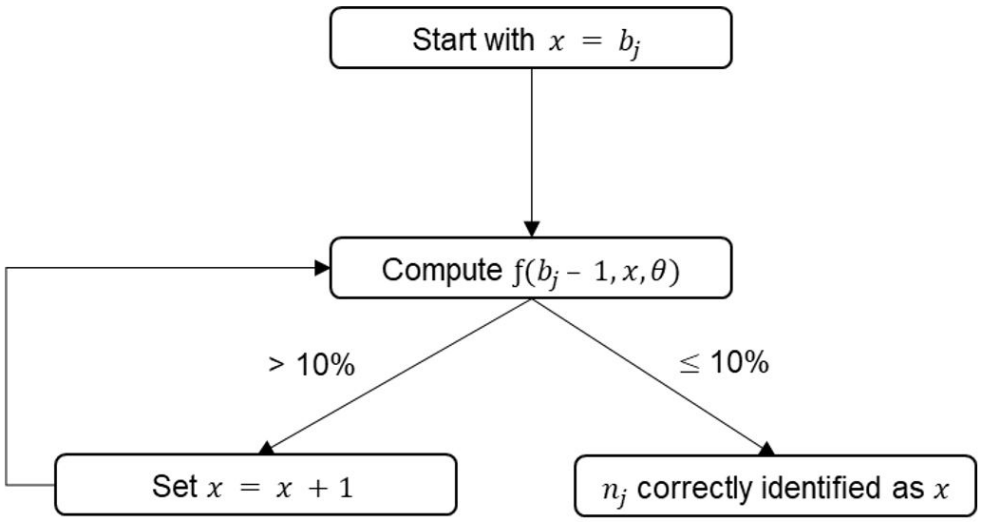


Figure 2.2. Flow chart for determining the cohort size based on specified DLTs (X), under safety threshold θ.

2.2.4. Selection of the number of cohort sizes

The final element of the design is the number of cohort sizes, which we denote J. As a rule, for a fixed b_j, the smaller the value of θ, the larger will be the corresponding n_j. Therefore, for small values of θ it is preferable to choose a smaller J; otherwise, the total sample size of the cohort may be excessive. In the next subsection we demonstrate some feasible choices of θ and J.

2.2.5. Example designs

Flexibility of the design derives from the safety threshold θ. We note that one should think of θ as a tuning parameter and not a target DLT rate; in simulations (see Section 2.3 below), we show that θ typically exceeds the modal DLT rate by 10%–20%. For example, $\theta = 0.25$ leads to a low maximum DLT rate; $\theta = 0.35$ to a moderate rate; and $\theta = 0.50$ to a high rate. Thus a strategy for identifying a design is to simulate frequency properties under

a likely dose-response curve for a range of values of θ and J, selecting the pair that gives the desired target DLT rate with a feasible sample size.

We use the notation CS($100\theta; n_1, \ldots, n_J$) to denote a cohort-sequence design using θ as the toxicity threshold and $n_1, \ldots, n_J$ as the sequence of cohort sizes, assuming by default that $b = (1, 2, \ldots, J)$. For cohort-sequence designs with high maximum acceptable DLT rate we set $J = 5$ and $\theta = 0.5$, which leads to cohort sizes $n = (1, 3, 5, 8, 10)$. We denote this design CS(50;1,3,5,8,10), or henceforth CS(50). For moderate maximum acceptable DLT rate, we set $J = 2$ and $\theta = 0.35$ and derive $n = (2, 6)$, designated CS(35;2,6). For low maximum acceptable DLT rate, we set $\theta = 0.25$ and $J = 2$ and compute $n = (5, 11)$, designated CS(25;5,11). Alternatively, letting $\theta = 0.4$ and $b = (1, 2, 3)$, the cohort sizes are $n = (3, 6, 9)$, which is similar to the 3+3 with an expansion cohort of size 3; we designate this design CS(40;3,6,9). We present these n, b combinations in Table 2.1.

Table 2.1. Decision rules for cohort-sequence designs used in the simulation.

	Sequence of cohort sizes (n)			
DLT limit (b_j)	CS(25;5,11)	CS(35;2,6)	CS(40;3,6,9)	CS(50)
1	5	2	3	1
2	11	6	6	3
3			9	5
4				8
5				10

With $J > 5$ or $\theta < 0.3$, the possible total number of subjects in the study can exceed the sample sizes that are typical for phase I trials. On the other hand, with $J < 5$, one may not achieve the numbers of subjects at the candidate MTD that we typically observe in practice when θ is large. With $J = 5$, the cohort size at the identified MTD is similar to the sizes of typical dose-expansion cohorts, making it unnecessary to enroll additional patients at those doses.

We have implemented the cohort-sequence design in the R package *cohortsequence*, which computes $f(X_i, N_i, \theta)$ and can calculate b_j from n_j and *vice versa*. The package also provides a function to simulate the performance of a cohort-sequence design for a designated dose-response scenario.

2.3. Simulation

2.3.1. Design

Ahn conducted simulations to compare the 3+3 with variants of the model-based continual reassessment method (CRM) design. (2) We used the same dose/toxicity scenarios as in his paper, comparing cohort-sequence designs for $\theta \in \{0.25, 0.35, 0.4, 0.5\}$. Our simulations terminated cohorts as soon as the number of DLTs was high enough to signal toxicity, a practice that improves efficiency and eliminates inadmissible toxicity.

We compared the performance of seven dose-escalation designs:

1. The traditional 3+3.
2. The 3+3 with an expansion cohort of size 3, with the stopping bound for safety in the expansion cohort set to 2;(32) we denote this design as 3+3@9,2. With this design, we de-escalate from an unsafe expansion cohort and enroll another until the final dose is safe.
3. CS(40;3,6,9): A cohort-sequence design that is similar to the 3+3@9,2.
4. CS(25;5,11): Suitable for a low target DLT rate.
5. CS(35;2,6): Suitable for a moderate target DLT rate.
6. CS(50;1,3,5): Suitable for a higher target DLT rate with fewer patients.

7. CS(50): Suitable for a higher target DLT and including a built-in expansion cohort.

We repeated the simulation 5000 times. When the lowest dose was rejected as toxic, the estimated MTD was designated as dose level 0. In this case, the number of patients treated at the estimated MTD is 0, and we do not enroll an expansion cohort. We used various frequency measures to compare performance: The proportion of times each dose was recommended as the MTD, the fraction of patients treated at each dose, the average number of patients enrolled, and the average proportion of patients experiencing a DLT.

2.3.2. Results

Tables 2.2–2.5 display simulated frequency properties of the designs applied to Ahn's scenarios. When the target toxicity is $\leq$ 10%, the CS(25;5,11) design gives the correct estimate most frequently. When the target toxicity is between 10% and 25%, the prediction accuracy for CS(35) is higher. When the target toxicity exceeds 25%, designs with $\theta = 0.5$ lead to correct estimates most often. Cohort-sequence designs generally treat lower fractions of subjects at low, safe doses. An exception is the CS(40;3,6,9), which closely mimics the behavior of 3+3.

The 3+3@9,2 and CS(40;3,6,9) designs require 4 to 6 more patients than the traditional 3+3, with the extra patients constituting a built-in dose-expansion cohort. For cohort-sequence designs with $\theta = 0.5$, the average number of patients with $J = 3$ is smaller than with $J = 5$, although other frequency properties are similar. The CS(50;1,3,5) in particular requires fewer patients than the 3+3. CS(35;2,6) performs similarly to 3+3 in terms of MTD recommendation and patient allocation, but it requires 3 fewer patients in toxicity scenario 4 and 1 fewer patient in the other scenarios.

With a higher value of the tuning parameter θ, the realized toxicity fraction is typically higher. Yet even for the CS(50) designs, which deliberately target higher toxicity rates,

the fractions of subjects experiencing toxicity are less than 35% under all 4 scenarios. The toxicity percentages for the traditional 3+3 design are similar to those for CS(35;2,6) and CS(40;3,6,9).

CS(50) designs, with their greater tolerance for DLTs, assign more patients at higher dose levels. Nevertheless, they effectively avoid extremely toxic doses, as subjects rarely reach dose levels with DLT rates in excess of 50%. The CS(25;5,11) design enrolls subjects at these highly toxic doses only in scenarios 3 and 4, where there is a steep jump from 25% DLTs to 80% DLTs in one dose elevation. Even so, it enrolls fewer subjects at those levels than the 3+3 and 3+3-like cohort-sequence designs.

2.4. Application

It is generally impossible to compare designs on a "live" data set, because any real data would have arisen under a design that dictated a sequence of dose assignments that another design would not replicate. To attempt a realistic comparison of designs, we generated DLT responses using a probit model estimated from the data of Simon *et al.*(46) The model assumes $Y_i = \log(d_i) + \epsilon_i$ with $\epsilon_i \sim \mathbb{N}(0, \sigma^2)$, and registers a DLT if $Y_i > K$. We estimated the parameters to be $\sigma = 1.092$ and $K = 8.78$, which lead to DLT probabilities of 1%, 5%, 14%, 32%, and 56% at doses 11, 13, 15, 17, and 19, respectively. For each subject, we generated a normal error ϵ_i and created a corresponding latent outcome under each dose d_i. In this way we created an ensemble of correlated data sets, one for each design.

Results appear in Table 2.6. All the methods estimate the MTD as either dose 15 (DLT rate 14%) or dose 17 (DLT rate 32%). The 3+3 with no expansion cohort yields a moderate sample size, a small fraction of DLTs, and the lowest proportion of subjects at or near the ultimate MTD. The 3+3@9,2 and the similar CS(40;3,6,9) give identical results; compared to the 3+3, they have larger sample sizes, comparable fractions of DLTs, and larger proportions of subjects treated at or near the MTD. CS(35;2,6) and CS(50) give equal

or larger proportions of DLTs than the 3+3-type designs, but also treat more patients at or near the estimated MTD. Notably, both CS(50) and CS(50;1,3,5) enroll fewer subjects and estimate the MTD as dose 17. CS(40;3,6,9) has a low DLT proportion comparable to that of the 3+3 design, but treats a larger fraction of subjects near the estimated MTD. These advantages come at the expense of a larger sample size.

2.5. DISCUSSION

We have proposed a family of rule-based phase I designs that retains the simplicity of the 3+3 while addressing its inflexibility and inefficiency.

We note that we have examined our design only in conventional scenarios where toxicity rates increase — sometimes slowly, sometimes quickly — with increasing doses. Some contemporary cancer treatments, such as immunotherapies, have toxicity curves that are not strictly monotonic. In such a case, one may wish to use a method that aims to optimize response subject to a maximum acceptable toxicity. From the standpoint of a statistician tasked with designing a trial whose main objective is to study *safety*, the key consideration must be to identify, and pull back from, toxic doses. We therefore believe that there is a continuing role for traditional designs that operate on this principle, even in trials where there is a strong prior expectation of a non-monotone toxicity curve.

Unlike the 3+3, which targets DLT rates in the range 20%–25%, with our approach one can select a design to reflect any targeted DLT rate by means of the tuning parameter θ. Simulations suggest that choosing θ to be 10%–20% higher than the target toxicity probability gives the best chance of having the target dose be the modal dose, although this varies by scenario. The choice of J, the maximum number of cohort sizes, largely controls the total number of patients enrolled. If the sample size available for the trial is comparable to those typically in use with 3+3 designs, then $J = 2$ works well for lower target toxicity rates, whereas $J \in \{3, 4, 5\}$ works well for target toxicity rates of 25% or higher. When

the target toxicity is between 10% and 25%, CS(35;2,6) is a practical choice. When the target toxicity exceeds 25%, we recommend CS(50;1,3,5) for a smaller total sample size or CS(50;1,3,5,8,10) when more patients are available. Using a large final cohort increases sample size requirements but eliminates the need for an add-on dose-expansion cohort.

Our cohort-sequence design improves efficiency by escalating rapidly through the lower, safer doses and increasing cohort sizes adaptively when one begins to encounter toxicities. Consequently, it generally enrolls more patients in the vicinity of the estimated MTD and incurs higher overall DLT rates. This is an advantage of the cohort-sequence approach, which avoids the wasteful assignment of subjects to doses that are likely to be safe and ineffective. In this way the cohort-sequence design paints a clearer picture of the drug's toxicity profile and increases the chance of clinical responses.

Our design effectively generalizes the 3+3; the CS(40;3,6,9) version is comparable to the 3+3@9,2, which is a 3+3 with added dose-expansion cohort. Unsurprisingly, these two designs and the 3+3 perform comparably on most metrics, except that the 3+3 enrolls fewer patients because it lacks a built-in expansion cohort. The CS(35;2,6) also performs similarly to the 3+3, with slightly fewer patients.

Although the critical values of the cohort-sequence design reflect Bayesian notions of parameter uncertainty, unlike the model-based designs it estimates the MTD based only on information from patients treated at the identified MTD. An alternative, hybrid approach that uses all the data would be to run the study with a rule-based design and then estimate a model (such as the logistic) for the dose-response data, designating as MTD the dose whose predicted DLT rate is closest to, but does not exceed, the target rate.

Model-based designs such as the CRM aim to identify the dose that delivers a targeted DLT rate. (39) By assuming an underlying dose-response model, all the data come into play at every decision point. These designs are both accurate and efficient, provided only that the assumed model is roughly correct. Their shortcomings are twofold: First, there is the

need to specify a model based on little or no prior data, and therefore some dependence on the assumed model and the prior distribution is inevitable. Second, the conduct of such studies requires substantial attention, potentially a complete reanalysis of the data and re-evaluation of the dosing scheme at the time of enrollment of each new patient. (34) This activity requires more statistician and clinician time than many institutions can afford to allocate. This appears to be the primary reason that these designs have not come into wide use, in spite of their well-documented excellent statistical properties.

Table 2.2. Comparison of phase I designs under dose-toxicity scenario 1.

		Designs						
Dose	DLT rate	3+3	3+3@9,2	CS(40;3,6,9)	CS(25;5,11)	CS(35;2,6)	CS(50;1,3,5)	CS(50)
Proportion of recommended dose (MTD) dose levels (%)								
0		3	4	3	9	2	0	0
1	0.05	10	17	10	26	9	2	2
2	0.10	38	50	38	51	36	12	12
3	0.25	31	22	32	13	33	19	23
4	0.35	16	6	15	1	17	32	39
5	0.50	2	0	2	0	3	28	22
6	0.70	0	0	0	0	0	6	2
7	0.80	0	0	0	0	0	1	0
Proportion of patients treated (%)								
1	0.05	27	26	24	41	21	10	7
2	0.10	31	34	32	38	29	15	14
3	0.25	25	25	27	18	28	20	20
4	0.35	13	11	14	3	16	23	27
5	0.50	4	3	4	0	5	20	22
6	0.70	0	0	0	0	1	9	8
7	0.80	0	0	0	0	0	2	1
Average number of patients		15.3	19.9	19.6	21.3	14.5	12.8	20.0
Percent toxicity (%)		20.3	19.1	18.6	13.3	20.7	34.7	34.3

Table 2.3. Comparison of phase I designs under dose-toxicity scenario 2.

		Designs						
Dose	DLT rate	3+3	3+3@9,2	CS(40;3,6,9)	CS(25;5,11)	CS(35;2,6)	CS(50;1,3,5)	CS(50)
Proportion of recommended dose (MTD) dose levels (%)								
0		3	3	3	9	2	0	0
1	0.05	10	13	9	24	8	2	2
2	0.10	18	22	16	31	14	4	4
3	0.15	21	24	19	23	22	6	6
4	0.20	20	21	21	10	22	9	10
5	0.25	18	13	20	3	19	17	21
6	0.35	9	4	11	0	11	31	36
7	0.50	2	0	1	0	2	26	20
8	0.75	0	0	0	0	0	4	1
9	0.90	0	0	0	0	0	0	0
Proportion of patients treated (%)								
1	0.05	24	23	21	38	18	9	7
2	0.10	23	23	22	31	21	10	8
3	0.15	20	21	20	20	21	12	10
4	0.20	15	16	17	8	18	13	12
5	0.25	10	10	12	3	13	15	17
6	0.35	5	5	6	0	7	17	22
7	0.50	2	1	2	0	2	15	17
8	0.75	0	0	0	0	0	6	5
9	0.90	0	0	0	0	0	1	1
Average number of patients		18.7	23.3	24.9	24.6	17.7	15.6	22.6
Percent toxicity (%)		18.1	17.5	17.2	12.9	18.4	30.5	31.3

Table 2.4. Comparison of phase I designs under dose-toxicity scenario 3.

		Designs						
Dose	DLT rate	3+3	3+3@9,2	CS(40;3,6,9)	CS(25;5,11)	CS(35;2,6)	CS(50;1,3,5)	CS(50)
Proportion of recommended dose (MTD) levels (%)								
0		10	11	9	26	7	0	2
1	0.10	9	9	8	17	7	2	2
2	0.10	8	9	7	15	7	2	2
3	0.10	7	12	6	12	8	1	1
4	0.10	30	39	31	23	32	13	13
5	0.25	36	20	39	6	38	76	80
6	0.80	0	0	0	0	0	4	0
7	0.90	0	0	0	0	0	0	0
Proportion of patients treated (%)								
1	0.10	29	28	26	47	23	11	9
2	0.10	19	18	17	22	18	10	8
3	0.10	16	15	15	14	16	11	8
4	0.10	16	20	19	11	19	15	14
5	0.25	15	16	19	5	19	33	46
6	0.80	4	4	4	0	5	18	14
7	0.90	0	0	0	0	0	2	2
Average number of patients		18.2	22	23.5	24	17.2	13.3	19.0
Percent toxicity (%)		20	19.4	18.7	17.5	19.2	30.1	29.1

Table 2.5. Comparison of phase I designs under dose-toxicity scenario 4.

		Designs						
Dose	DLT rate	3+3	3+3@9,2	CS(40;3,6,9)	CS(25;5,11)	CS(35;2,6)	CS(50;1,3,5)	CS(50)
Proportion of recommended dose (MTD) dose levels (%)								
0		0	0	0	0	0	0	0
1	0.01	0	0	0	0	0	0	0
2	0.01	3	4	3	8	2	1	1
3	0.05	10	17	9	26	10	2	2
4	0.10	41	52	41	52	40	15	15
5	0.25	46	26	47	13	48	79	82
6	0.80	0	0	0	0	0	4	0
7	0.90	0	0	0	0	0	0	0
Proportion of patients treated (%)								
1	0.01	16	13	13	17	12	8	6
2	0.01	16	14	15	19	13	9	6
3	0.05	19	19	17	24	17	10	7
4	0.10	22	27	25	26	24	15	15
5	0.25	21	22	25	13	26	79	50
6	0.80	6	5	4	1	7	4	15
7	0.90	0	0	0	0	0	0	2
Average number of patients		20	24.3	23.9	32.2	17.1	12.5	18.3
Percent toxicity (%)		14.4	14	13.2	8	15.2	29.0	28.1

Table 2.6. Application of the designs to data generated according to the trial in Simon *et al.*(46) CS(25) represents CS(25;5,11), and CS(35) represents CS(35;2,6).

	Design						
	3+3	3+3@9,2	CS(40;3,6,9)	CS(25)	CS(35)	CS(50;1,3,5)	CS(50)
Patients enrolled	17	20	20	27	13	11	15
DLT rate at $\widehat{\text{MTD}}$ (%)	32	32	32	14	14	32	32
% with a DLT	18	15	15	11	23	18	40
% treated at $\widehat{\text{MTD}}$	35	45	45	41	46	45	67
% treated within 1 level of $\widehat{\text{MTD}}$	65	70	70	81	85	82	87

CHAPTER 3

PROJECTING CLINICAL TRIAL RESULTS TO A TARGET POPULATION

3.1. Introduction

A clinical trial lacks *external validity* if the sample in which it takes place differs in important ways from the target patient population in which one intends to apply its findings.(7, 33, 52) Of particular concern is the estimated treatment effect, which describes the potency of a new treatment and predicts the consequences of its widespread adoption. A trial whose estimated treatment effect differs from its value in the target population may misinform clinical practice, leading to sub-optimal therapy of future patients.

If one possesses data on the distribution of predictors of outcome in the target population, and one expects treatments to have similar effects in relevant subsets of the trial and target populations, then it is in principle possible to project treatment effects to this population.(27, 40, 53) We denote this process *interpolation*, because we assume that all relevant subsets occur in both the trial sample and the target population.

In the traditional interpolation method of *post-stratification*, we partition the trial participants into strata defined by a set of covariates that exhibit interaction with treatment effects. We compute treatment effects within these strata, and average the effects weighted by the fractions of subjects in the population that lie within the relevant strata. The method is model-independent, and is known to work well when the list of confounders is short.(48)

Contemporary approaches to interpolation rely on variants of propensity score analysis, in which one estimates the probability of trial inclusion for each subject in the trial. One can then use the inverse inclusion probabilities as weights in constructing a corrected population treatment effect estimate.(10, 49) Stuart *et al.* applied such a method to data on a school-based cluster trial that was conducted in Maryland, using a list that contained relevant demographic data for all eligible schools in that state.(48)

Alternatively, one can estimate the population treatment effect by creating strata based on the inclusion probabilities, analogous to a stratified propensity score analysis. One then constructs stratum-specific treatment effect estimates and averages them to obtain an overall estimate that represents the target population. This approach avoids the problem of excessively variable weights and therefore can be numerically more stable.(38) Matching groups using trial inclusion probabilities is similar to dividing the subjects into subgroups when the two methods use the same classification variables.(49, 50)

Inclusion probability methods are rarely applicable in medical trials, because the necessary population data set — a census of trial-eligible patients with all relevant confounders recorded — seldom exists. Potential exceptions include rare diseases with comprehensive registries, or diseases that are recorded in national health databases, as some European countries maintain. More commonly, the best population information would come from health surveys such as the US National Health & Nutrition Examination Survey (NHANES). (8) But although NHANES collects extensive demographic, behavioral, and health history data, its disease variables are largely self-reported and therefore potentially imprecise. One may be able to use NHANES as a reliable basis of population data for trials of common, non-specific conditions like hypertension and hypercholesterolemia, but it is less useful for conditions whose diagnosis depends on elaborate or expensive tests or procedures.

In this article we discuss these methods and illustrate their application in pharmaceutical research. We begin by describing a clinical trial that manifests a need for interpolation.

3.2. Example: The LRC-CPPT

The Lipids Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) was a randomized, double-masked, placebo-controlled study of the effect of cholestyramine on long-term cholesterol reduction and incidence of coronary heart disease.(30) Briefly, the study randomized over 3,700 men within strata defined by cardiovascular risk. Investigators directed participants to consume six packets a day of the assigned treatment — either active cholestyramine or a placebo. Follow-up visits took place two weeks after randomization and at 2-month intervals for an average of 7.4 years. The outcome variable is the difference between average cholesterol measurements taken after randomization and a weighted average of baseline measurements.(14)

Several measured covariates in the LRC-CPPT trial have different distributions in other populations where one might wish to treat for hypercholesterolemia. For example, the proportion of ever-smokers in the study, conducted between 1973 and 1983, was 76%, whereas in the US population currently it is 45%.(8) If smoking moderates cholestyramine effects, an estimated treatment effect derived only from trial data may be misleading.

3.3. Methods

3.3.1. Estimating the treatment effect in the trial

Let $Y_i(Z_i)$ be the potential outcome for subject i when randomized to treatment Z_i ($1 =$ experimental, $0 =$ control). The causal effect of randomization for subject i is the difference between its experimental and control potential outcomes: $Y_i(1) - Y_i(0)$. The causal effect in the trial population is the expectation of the average difference in individual causal effects

in the trial

$$\Upsilon = \mathbb{E}_{\mathcal{R}}\left[\frac{\sum_{i=1}^{n} Y_i(1) - Y_i(0)}{n}\right],$$

where n is the trial sample size and $\mathbb{E}_{\mathcal{R}}$ refers to the sampling of subjects into the trial.(21)

Because we can observe a subject's outcome only under the assigned treatment, we denote the observed outcome $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$. Under assumptions of random treatment assignment and no interference between units (the *stable unit treatment value assumption*), one can compute an unbiased estimate of the causal effect of Z on Y as

$$\hat{\Upsilon} = \frac{\sum_i Z_i Y_i}{\sum_i Z_i} - \frac{\sum_i (1 - Z_i) Y_i}{\sum_i (1 - Z_i)}, \tag{3.1}$$

also known as the *intention-to-treat (ITT)* estimate of the treatment effect. Its standard error (SE) is

$$\mathrm{SE}(\hat{\Upsilon}) = \sqrt{\frac{s^2_{Y,1}}{\sum_i Z_i} + \frac{s^2_{Y,0}}{\sum_i (1 - Z_i)}},$$

where $s^2_{Y,0}$, $s^2_{Y,1}$ are the sample variances of Y in the groups randomized to control and experimental treatment, respectively. Alternatively, one can estimate Υ *via* a linear regression model on Y with an intercept, a slope on Z, and normal residuals E:

$$Y = \alpha + Z\Upsilon + E. \tag{3.2}$$

This approach gives the same result as estimating the ITT effect of Z on Y in Equation (3.1), with the difference that the SE is based on a pooled variance estimate.

Another analysis — denoted an *analysis of covariance* — would estimate the treatment effect with model-based adjustment for baseline covariates X:

$$Y = \alpha + X\beta + Z\Upsilon + E. \tag{3.3}$$

In a randomized trial, the estimate of Υ from Model (3.3) will differ from that from Model (3.2) due to variation in the distribution of X, but their expectation is the same. To detect whether there is *heterogeneity* of the treatment effect — defined here as meaning that the treatment effect varies across subgroups defined by the covariates — one can create a regression model that includes the randomization indicator, covariates, and interactions:

$$Y = \alpha^\dagger + X\beta^\dagger + Z\Upsilon^\dagger + ZX\gamma^\dagger + E. \tag{3.4}$$

We refer to the estimand here as $\Upsilon^\dagger$, because it is not in general the same parameter as Υ, which represents a treatment effect averaged over the distribution of covariates in the trial. We call a covariate that manifests a significant interaction with the randomization variable a *moderator* of the treatment effect. Note that the error variances estimated from Models (3.2), (3.3), and (3.4) differ in expectation, generally decreasing as one adds terms to the model.

3.3.2. Interpolation with post-stratification

3.3.2.1. Post-stratification with a discrete stratifier

The objective of our analysis is to project a standard clinical trial treatment effect estimate to its likely value in the ultimate target population. For notational convenience, we recode the moderator X variables into a single stratifier S that takes values s_j, $j = 1, \ldots, J$. In medical trials the stratifiers would typically be some combination of demographic variables, pre-treatment risk factors such as smoking status, indicators of disease type or severity, and biomarkers. We assume that we know the distribution of S in the target population.

Let $f^{\mathcal{T}}(s)$ and $f^{\mathcal{P}}(s)$ represent the probability mass functions of the stratifier in the trial and target population, respectively. Letting $\hat{\Upsilon}_j$ be the estimated treatment effect from

stratum j, then the raw clinical trial estimate $\hat{\Upsilon}$ from Equation (3.1) is approximately

$$\hat{\Upsilon} \approx \hat{\Upsilon}^{\mathcal{T}} = \sum_{j=1}^{J} f^{\mathcal{T}}(s_j)\hat{\Upsilon}_j.$$

The two estimates are only approximately equal because slight imbalances will arise in the randomization within strata in finite samples.

To create an estimated treatment effect $\hat{\Upsilon}^{\mathcal{P}}$ that reflects the distribution of the stratifier in the target population, one simply re-weights the stratum-specific estimates:

$$\hat{\Upsilon}^{\mathcal{P}} = \sum_{j=1}^{J} f^{\mathcal{P}}(s_j)\hat{\Upsilon}_j. \tag{3.5}$$

By independence across strata, the standard error of (3.5) is

$$\mathrm{SE}(\hat{\Upsilon}^{\mathcal{P}}) = \sqrt{\sum_{j=1}^{J} \left[f^{\mathcal{P}}(s_j)\right]^2 \mathrm{SE}^2\left(\hat{\Upsilon}_j\right)}. \tag{3.6}$$

This is the classical post-stratification estimate and variance.

It is elementary to show that if $\mathbb{E}[\hat{\Upsilon}^{\mathcal{T}}] \neq \mathbb{E}[\hat{\Upsilon}^{\mathcal{P}}]$, then i) there is a treatment-by-stratum interaction; i.e., $\exists\ j \neq k$ such that $\mathbb{E}[\hat{\Upsilon}_j] \neq \mathbb{E}[\hat{\Upsilon}_k]$; and ii) the trial and population weights differ; i.e., $\exists\ j : f^{\mathcal{P}}(s_j) \neq f^{\mathcal{T}}(s_j)$. Clearly, if treatment effects do not vary across strata, then the weighting scheme cannot alter the expectation of the weighted estimate. Moreover, unless the trial and population stratum proportions differ, again reweighting can have no effect. Although these conditions are not sufficient, a trial in which both are satisfied presents a *prima facie* case for interpolation.

One can implement post-stratification using the `ESTIMATE` statement in `SAS PROC GLM`. For example, if smoking is the sole stratifier, one can estimate the effect among smokers using "`ESTIMATE Z 1 -1 Z*smoke 1 0 -1 0/E;`" and the effect among non-smokers using "`ESTIMATE Z 1 -1 Z*smoke 0 1 0 -1/E;`". To post-stratify by smoking, where the pro-

portion of smokers in the population is, say, 80%, one can use "`ESTIMATE Z 1 -1 Z*smoke 0.8 0.2 -0.8 -0.2/E`".

3.3.2.2. Post-stratification with a continuous stratifier

When S is so fine as to be effectively continuous, each stratum may contain only one or a few observations, and thus estimates can be undefined within strata. Instead, we use a smoothing approach. For each s_j (assuming sorted values), create the moving window set

$$\sigma_j = \{\max(s_j - d, \min(S)), \ldots, \min(s_j + d, \max(S))\}. \tag{3.7}$$

and replace $\hat{\Upsilon}_j$ in (3.5) by the corresponding treatment effects estimated on σ_j. An alternative smoother computes stratum treatment effects using the k nearest neighbors. That is, for each s_j, include in σ_j the k subjects whose S values are closest to it.

When multiple subjects share the same S value, the resulting subset can have more than k subjects. In LRC-CPPT we implemented this method, interpolating on age with $k \approx 5N/J$. To avoid over-smoothing on the boundaries, we set $k \approx 3N/J$ when $j = 1$ or $j = J$; in our applications, this gave similar results to a window size of $d = 2$. Estimates on the boundaries are typically smoother with the nearest-neighbors approach. With multiple continuous stratifiers, one can define a multivariate measure for distance, then extract the k nearest neighbors.

For continuous S, we generally do not have independence across strata, and therefore we cannot use Equation (3.6) to compute standard errors. Instead, we compute a bootstrap SE under strategy (3.7), resampling (Z_i, Y_i) within strata.

When there are multiple stratifiers, all of them continuous, one can select discrete pseudo-strata for each point based on nearest neighbors according to some multivariate distance measure. When there are both continuous and discrete stratifiers, it is preferable to first

select pseudo-strata by continuous predictors, and then stratify within those based on discrete stratifiers. This avoids the problem of empty strata, as occurs in LRC-CPPT, where there are, for example, no non-smokers aged 34.

3.3.3. Interpolation with inverse inclusion probability weighting

Stuart *et al.* (48) proposed computing for each subject the probability of being selected into the trial, then using the inverse of the inclusion probability, W_i^{IIPW}, as a weight in estimating causal effects. The estimate of the effect of Z on Y using inverse inclusion probability weighting (IIPW) is

$$\hat{\Upsilon}^{\mathcal{P},\text{IIPW}} = \frac{\sum_i Z_i Y_i W_i^{\text{IIPW}}}{\sum_i Z_i W_i^{\text{IIPW}}} - \frac{\sum_i (1-Z_i) Y_i W_i^{\text{IIPW}}}{\sum_i (1-Z_i) W_i^{\text{IIPW}}}. \tag{3.8}$$

When the stratification is categorical and includes all covariates, the trial inclusion score for subject i in stratum j is $p_{ij} = \frac{n_j}{N_j}$, leading to a weight (now indexed by stratum and subject)

$$W_{ij}^{\text{IIPW}} = \frac{1}{p_{ij}} = \frac{N_j}{n_j},$$

where n_j and N_j represent the total numbers of subjects in stratum j in the trial and target populations, respectively. In the Appendix, we show that the individual inverse inclusion probability weights approximate the weights implied in the post-stratification scheme using the same set of stratifiers. The two approaches are equivalent when randomization fractions are equal in all strata, as would occur in large samples.

3.3.4. Interpolation with inclusion probability stratification

One can also use the trial inclusion probabilities in an alternative post-stratification approach, we denote this as inclusion probability stratification (IPS). Specifically, one re-

defines S by creating pseudo-strata that group together the trial observations based on their trial inclusion scores. One then computes treatment effects within the strata and use Equation (3.5) to reweight them based on the distribution of S in the target population. As with propensity score analysis, dividing into five strata of equal size appears to be sufficient to reduce the bias to negligible levels.(42) Moreover, stratification with a modest value of J leads to an estimate that is more numerically stable than $\hat{\Upsilon}^{\mathcal{P},\text{IIPW}}$.

3.3.5. Computing standard errors with inclusion probabilities

3.3.5.1. The standard error of the IIPW estimate

One can compute the IIPW estimate either directly using Equation (3.8) or through a weighted regression. Setting $U = [1\ Z]$ and defining W to be a square matrix with the weights W_i on the diagonal, one obtains

$$\hat{\delta}^{\mathcal{P},\text{IIPW}} = (U^{\text{T}}WU)^{-1}U^{\text{T}}WY \tag{3.9}$$

where $\hat{\delta}^{\mathcal{P},\text{IIPW}} = [\hat{\alpha}^{\mathcal{P},\text{IIPW}},\ \hat{\Upsilon}^{\mathcal{P},\text{IIPW}}]$. The estimated variance is

$$\mathbb{V}(\hat{\delta}^{\mathcal{P},\text{IIPW}}) = (U^{\text{T}}WU)^{-1}U^{\text{T}}W\hat{\sigma}^2IWU(U^{\text{T}}WU), \tag{3.10}$$

where $\hat{\sigma}^2$ is the sample variance of the residuals, divided by W. Although $\hat{\Upsilon}^{\mathcal{P},\text{IIPW}}$ is consistent for the population treatment effect $\Upsilon^{\mathcal{P}}$, its nominal residual variance may greatly exceed that created from classical post-stratification. This is because the latter is based implicitly on a model that includes all relevant predictors of outcome (as in Model (3.4)), whereas the former includes only Z (as in Model (3.2)). To correctly estimate the residual variance and treatment effect variance, one should substitute the estimated residual variance from Model

(3.4) into Equation (3.10).

3.3.5.2. The standard error of the IPS estimate

Similarly, the IPS approach gives an inflated SE if there is substantial heterogeneity in treatment effects. Here we use the post-stratification approach with an SE computed within each stratum:

$$\mathrm{SE}(\hat{\Upsilon}_j) = \hat{\sigma}\sqrt{\frac{1}{n_{j1}} + \frac{1}{n_{j0}}}, \tag{3.11}$$

with n_{j1} and n_{j0} equal to the number of patients in stratum j in the experimental and control arms, respectively. We apply Equation (3.6) with $f^{\mathcal{P}}(s_j)$ to get the correct SE. Again, one should substitute the residual standard deviation from Model (3.4) into Equation (3.11).

By contrast, if in the conventional post-stratified analysis we assume a common value of the within-stratum variance σ^2, then this equals the variance under Model (3.4), and there is no need to further adjust the standard errors.

3.4. Simulation study

We conducted a one-replicate simulation study to illustrate the potential bias from omitting important covariates in projecting a population treatment effect estimate. We created a synthetic population consisting of six copies of the LRC-CPPT trial dataset (randomization and covariates only). Next we generated an outcome for which education, smoking, and age group exhibit no, moderate, and strong interaction, respectively, with treatment assignment:

$$
\begin{aligned}
Y \quad = \quad & -2 + 20 \times Z - 2 \times \texttt{smoke} + 3 \times \texttt{agegroup} \\
& +0.5 \times \texttt{education} - 0.25 \times Z \times \texttt{smoke} \\
& +12 \times Z \times \texttt{agegroup} + E,
\end{aligned}
$$

where E is a vector of independent $\mathbb{N}(0, 3^2)$ deviates. The variable `agegroup` is coded as equally spaced groups representing three age levels: < 44, 44–50, and ≥ 51. We also divided `education` into three groups — less than high school, high school graduate, and some college and above. Because there are substantial interactions, the population treatment effect, representing an average of treatment effects across strata, need not equal 20, the coefficient of the randomization indicator Z in the mean model. Finally, we selected a sample of subjects to constitute the trial population *via* the logistic model

$$
\text{logit}(p) = -2.5 - 0.5 \times \texttt{smoke} + 0.28 \times \texttt{age group} + 0.6 \times \texttt{education};
$$

here p is the probability of being selected into the trial. From the resulting synthetic population of 22,704 subjects we sampled 3,786 to become the synthetic trial dataset.

Table 3.1 displays estimated treatment effects from the synthetic population, the synthetic trial, IIPW using the true and estimated trial inclusion scores, IPS on inclusion scores, and post-stratification using all possible combinations of the three stratifiers. Our ground truth, from the synthetic target population, is $\Upsilon^{\mathcal{P}} = 32.11$; compare this to $\hat{\Upsilon}^{\mathcal{T}} = 33.58$ (SE=0.30) from the synthetic trial. Estimates from IIPW matched the target population values well, regardless of whether we used true or (correctly) estimated probabilities. The treatment effect from the stratified inclusion score analysis gave $\hat{\Upsilon}^{\mathcal{P}} = 31.42$ (SE =0.25), and therefore was off target by almost 3 standard errors; moreover its estimated standard error was twice that from a weighted analysis. Using the corrected residual variance estimate from

Model (3.4) — including treatment, strata, and interactions — we obtained an SE of 0.11.

The correlation between the post-stratification and the true trial inclusion score weights for our simulated data was 0.95; the relationship was not perfectly linear due to a slight lack of balance. Estimates by post-stratification were close to the truth and precise provided that age, the variable with the strongest interaction with treatment, appeared as a stratifier.

Table 3.1. Estimated treatment effect (SE) on drop in cholesterol (mg/dL) from simulated LRC-CPPT data. Selection into the trial was based on age group, smoking status, and education.

Method	$\hat{\Upsilon}^{\mathcal{T}}$
Synthetic trial (N = 3,786)	33.58 (0.30)
	$\Upsilon^{\mathcal{P}}$
Synthetic population (N = 22,704)	32.11
Trial inclusion score analysis	$\hat{\Upsilon}^{\mathcal{P}}$
IIPW with true score	32.22 (0.30)
IIPW with true score weighting, corrected variance	32.22 (0.10)
IIPW with estimated score	32.15 (0.30)
IIPW with estimated score, corrected variance	32.15 (0.10)
IPS analysis, corrected variance	31.42 (0.11)
Post-stratification	$\hat{\Upsilon}^{\mathcal{P}}$
Education	34.01 (0.30)
Smoking	33.54 (0.30)
Age group	32.23 (0.11)
Education+smoking	33.92 (0.31)
Education+age group	32.27 (0.11)
Smoking+age group	32.20 (0.10)
Education+smoking+age group	32.23 (0.11)

3.5. Analysis of the LRC-CPPT Data

A stepwise variable selection found age and smoking status to have significant interactions with treatment assignment (Table 3.2). Estimated treatment effects by smoking status and by race for LRC-CPPT appear in Table 3.3. The proportion of smokers in LRC-CPPT was 76%, compared to current estimates from NHANES that 45% of Americans are ever-smokers.(8) Applying Equation (3.5) with the distribution of smokers from NHANES we obtain $\hat{\Upsilon}^{\mathcal{P}} = 28.02$ (SE=0.90), compared to $\hat{\Upsilon}^{\mathcal{T}} = 26.91$ (SE=0.80) in the trial. Although the treatment effect appears to differ by race (Table 3.3), only 3.4% of trial subjects identified as black, and the race-treatment interaction was not significant ($P = 0.51$). Therefore, we excluded race as a stratifier in further analyses.

Table 3.2. Estimated coefficients (SE) from stepwise selection.

Intercept	Treatment	Age	Smoke	Treatment × Age	Treatment × Smoke
−9.30 (4.23)	7.11 (5.95)	0.35 (0.08)	0.24 (1.31)	0.51 (0.12)	−5.52 (1.84)

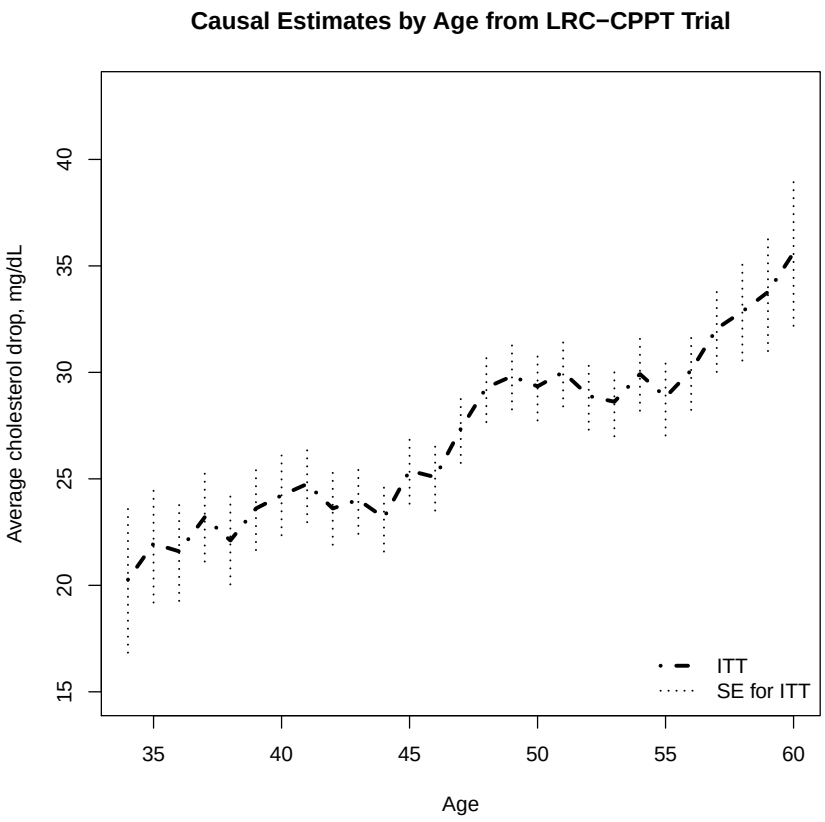
Table 3.3. Intention-to-treat estimates (SE) by stratum from LRC-CPPT; outcome is the drop in cholesterol (mg/dL).

	Smoking		Race	
	Non-smoker ($N = 904$)	Smoker ($N = 2880$)	Non-black ($N = 3655$)	Black ($N = 129$)
$\hat{\Upsilon}$	30.96 (1.66)	25.62 (0.91)	26.81 (0.81)	29.70 (4.63)

The age-by-randomization interaction was significant at $P < 0.0001$. As age is effectively continuous, we created interpolated estimates using moving windows of size $d = 2$ from ages 34 to 60. Figure 3.1 shows that the treatment effect increases dramatically with increasing age.

To obtain a population that better reflects disease status, we abstracted information from NHANES, which collects serum cholesterol values and therefore can identify subjects who

Figure 3.1. Estimates of average cholesterol drop by age.



would be eligible for treatment with a cholesterol-lowering drug. We combined data from 1999 to 2016, restricting attention to males aged 34–60.(8) Retaining only those who meet the LRC-CPPT hypercholesterolemia criteria, we ended up with 120 subjects. Expanding to the contemporary definition (total cholesterol $\geq$ 200 mg/dL; LDL cholesterol $\geq$ 100 mg/dL; triglycerides $\leq$ 300 mg/dL) yielded 2,118 subjects. From these data, we interpolated causal estimates by smoking and age, computing SE by the bootstrap. The interpolation led to a two mg/dL increase in the estimated treatment effect, representing approximately 2.5 standard errors; see Table 3.4.

We computed trial inclusion probabilities using hypercholesterolemic male NHANES participants to represent the target population. Eligible individuals were weighted by their NHANES sampling weights, which reflect their representation in the US population. Trial

subjects took weights of 1, as they represent only themselves. We estimated the inclusion probability with a logistic model including age and smoking status. We then conducted an analysis by inverse weighting of trial subjects' estimated inclusion probabilities, as well as an analysis stratified by trial inclusion probabilities. The results were similar to the interpolated estimates using measured stratifiers (Table 3.4) but 1 SE smaller than those from classical post-stratification. The SE for the interpolated estimates in the stratified analysis was similar to that using the pooled variance estimate from Model (3.4). The disparity was not as great as in the simulation study, because the interaction between age group and treatment was not as strong in the LRC-CPPT dataset.

Table 3.4. Interpolated treatment effect (SE) on drop in cholesterol (mg/dL) based on the distribution of smoking status and age in NHANES.

	$\hat{\Upsilon}^{\mathcal{T}}$
Trial	26.91 (0.80)

Analysis	$\hat{\Upsilon}^{\mathcal{P}}$
Post-stratification	28.92 (0.79)
IIPW	28.15 (0.79)
IPS	28.01 (0.87)

3.6. Discussion

We have described and applied methods for projecting a treatment effect estimate to a population that differs from the one in which the treatment was evaluated. Common assumptions of the methods are that i) subjects in the trial are similar to those in the target population, other than on a set of measured covariates, and ii) information is available on the distribution of these covariates in the population.

The drive to create personalized treatment strategies has heightened interest in the efficient estimation of treatment effects within subgroups. Nevertheless there is a continuing

need to estimate average treatment effects in the general population of subjects with a given disease. For example, economic evaluation, such as cost-effectiveness analysis, relies on estimated effects on survival and cost for the population. This is true whether the "treatment" is a single drug administered to all comers or a personalized strategy that depends on pharmacogenetic markers.

In our simulation, post-stratification on a covariate that displayed a strong interaction with treatment gave estimates that were close to the truth and to estimates from an inverse-probability-weighted approach that correctly used all relevant confounders. This suggests that interpolation by post-stratification can offer a reliable sensitivity analysis in the common situation where subject-level population data are unavailable. One can control the number of strata in the analysis by limiting attention to covariates that have strong interactions with treatment, as assessed from the trial data. We observed that an approach that segregates subjects into strata based on trial inclusion scores was less successful.

When all stratifiers are discrete, post-stratification weights are comparable to those based on inverse trial inclusion probabilities, and therefore the two methods give similar estimates. With continuous stratifiers, methods based on trial inclusion scores are potentially more accurate, because the scores can better reflect study inclusion probabilities. Stratum sizes may be small, however, and validity of estimates is hostage to correct model specification. Moreover, to obtain correct SEs with the inclusion probability approaches, one must conduct the extra step of estimating the variance with a model for Y that includes the treatment, the stratifiers, and treatment-stratifier interactions. When individual-level data are not available for the target population, it is not possible to directly compute trial inclusion probabilities. One can circumvent this problem by generating a pseudo-population using the estimated distribution of stratifiers. Post-stratification works well without the need for this extra step; moreover, it automatically avoids the problem of inflated SEs.

We followed the practice of the authors of the original example paper by taking the response variable to be the difference between follow-up and baseline values of cholesterol.(14)

This may be inefficient compared to estimation by analysis of covariance; (15, 45) in practice it is preferable to work with baseline-adjusted treatment effects within strata.

It is often said that trial participation has a positive effect, even for patients who are assigned to control therapy. This is because participants typically receive state-of-the-art treatment with frequent evaluation, and may enjoy other benefits such as reduced health care costs.(52) Patients who are likely to adhere well to treatment may also be over-represented in clinical trials. Nevertheless there is substantial evidence that treatment effects in randomized and observational studies are similar,(37, 40) supporting the use of interpolation as described here.

When the target population includes strata that are absent from the trial population, interpolation is impossible. LRC-CPPT, for example, excluded females and persons outside the age range 34–60. We refer to the projection of treatment effects to a population in which there are additional strata as *extrapolation*, because it involves averaging estimated stratum effects that lie outside the range of the trial data. For variables like age, one can extrapolate effect estimates using regression models, subject to the usual *caveat* about projecting beyond the range of the data. For variables like sex, where an entire stratum is absent by design, one can conduct sensitivity analyses, perhaps involving estimated treatment effects from other trials if any are available.

A related concern is that both the post-stratification and inverse-probability weighting assume that the clinical trial and target populations differ only with respect to the distribution of measured stratification variables. In some cases, the decision to participate in a trial may represent the outcome of a *nonignorable* selection mechanism, which one can think of as the target and trial populations differing also on unmeasured confounders. An important future direction is to address this concern *via* sensitivity analysis. (35)

3.7. Acknowledgements

We obtained the LRC-CPPT data from the NHLBI Biologic Specimen & Data Repository Information Coordinating Center. Our work does not necessarily reflect the opinions or views of the LRC-CPPT or the NHLBI.

CHAPTER 4

PROJECTING INSTRUMENTAL VARIABLE RESULTS TO A TARGET POPULATION

4.1. Introduction

A randomized clinical trial assesses the efficacy of a treatment among patients who participate in the trial, but its results may not be directly applicable to the target population for which the treatment is intended. Such concerns typically arise when the trial sample is not representative of the target population.(7, 33, 52) If the treatment effect differs by subgroups, and the subgroups are represented differently in the trial, then the treatment effect from the trial will differ from that in the target population. Such a trial may misinform clinical practice, leading to sub-optimal treatment of future patients.

If one can i) identify all covariates that have an interaction with treatment, ii) accurately measure the joint distribution of those covariates in the population, and iii) safely assume that the treatment effects in the target population match those in the trial, then it is in principle possible to project treatment effects from the trial to the target population. We denote this process as *interpolation*.

A classic method for interpolation is *post-stratification*, in which one reweights stratum treatment effects from the trial to the target population.(29) This approach requires only that one know the distribution of key covariates in the population. If individual-level data are available in the target population, it is possible to obtain estimated probabilities of trial inclusion for each subject in the trial. One can then obtain projections by using these

probabilities either in a weighting scheme (10, 48) or as the basis for post-stratification, applying a method reminiscent of propensity score analysis.(38)

The *intention-to-treat (ITT)* approach is the standard paradigm for clinical trial data analysis. The essence of this approach is to include all randomized subjects and to analyze a measured "hard" endpoint with subjects grouped according to randomization arm, even if some did not receive the assigned treatment. The estimand is the effect of the treatment as assigned, which is arguably more relevant to practice than its effect in a hypothetical situation of universal perfect adherence. We have discussed interpolation of ITT estimates previously.(29)

Alternatively, one may wish to estimate a treatment effect in *compliers*, or the notional population of subjects who would adhere to assigned treatment. A problem with this approach is that we cannot generally identify compliers. For example, some subjects who were assigned and received control therapy would also have taken control therapy if assigned the experimental intervention, in which case they are *never-takers* rather than compliers. Such considerations have led to the development of analysis methods grounded in the *Rubin causal model*.(21, 43, 44) This framework enables estimation of the treatment effect among compliers, the *complier average causal effect (CACE)*, by means of *instrumental variables* analysis.(3, 47) Other methods for examining treatment effects in light of compliance use *ad hoc* approaches (14) or more elaborate applications of the concept of principal stratification.(5, 16, 17, 26)

In this chapter we present a post-stratification approach that interpolates causal estimates from a clinical trial to reflect the distributions of predictive covariates in a larger, clinically relevant population using instrumental variable approach on CACE.(3) We compare post-stratification with two other approaches using trial inclusion probabilities. We assume that trial participants are similar — in terms of outcomes and compliance behavior — to nonparticipants other than on measured baseline factors.

4.2. Examples

4.2.1. LRC-CPPT

The Lipids Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) was a randomized, double-masked, placebo-controlled study of the effect of cholestyramine on long-term cholesterol reduction and incidence of coronary heart disease.(30) Over 3,700 men were randomized in strata grouped by cardiovascular risks. Participants took six packets of either cholestyramine or placebo each day (24 grams per day). Follow-up visits occurred two weeks after randomization and every two months, for an average of 7.4 years. Patients reported the number of assigned packets consumed during each follow-up visits. We take as outcome variable the difference between average cholesterol measurements taken after randomization and a weighted average of baseline measurements.(14)

Because cholestyramine was unpleasant to use, nonadherence was common. For the purposes of this analysis, we designate a complier as a patient who took at least 80% of his assigned packets, in which case roughly half of subjects in the cholestyramine arm are non-compliers. We seek to estimate treatment effects considering compliance.

LRC-CPPT participants differed from the contemporary population of US males both with respect to smoking habits and the distribution of age. We will use data from the trial and from the US National Health & Nutrition Examination Survey (NHANES) to examine the hypothesis that these differences could affect estimates of causal treatment effects.

4.2.2. New York School Choice Experiment

In 1997, the School Choice Scholarships Foundation conducted a study of the effect of private school vouchers on the educational achievement of children from low-income fami-

lies in New York City.(5, 18) The Foundation chose recipients by lottery from a cohort of applicant families. The control arm was a set of non-recipients who were pair-matched to the recipients *via* a propensity-score model. Thus treatment was assigned at random, albeit not in the conventional way. Subjects sat for the Iowa Test of Basic Skills, which evaluates ability in reading and math, both before randomization and after the following school year.

We take as outcome the grade-normed national centile rankings on reading and math. To avoid various complications, we focus on 715 subjects from single-child families with children in grades 1–4 during the application period and who had complete demographic, baseline, and follow-up data. Treatment adherence was imperfect, as some voucher recipients declined to attend private schools, and some non-recipients attended private schools at their own expense.

The cohort that we analyze was roughly equally divided between black and non-black students, whereas the population of NYC at the time was 30% black, and the US population was 14% black. As treatment effects plausibly differ between races, causal effects of the intervention, were it to be applied across all of NYC or the entire US, could differ substantially from those observed in the study.

4.3. Causal estimands and their estimates

4.3.1. The ITT estimand

Henceforth, we adopt the notation of the Rubin causal model.(3) Let $Y_i(Z_i, D_i)$ be the potential outcome for subject i, $i = 1, \ldots, n$, when randomized to treatment Z_i and taking treatment $D_i = D_i(Z_i)$ (1 = experimental, 0 = control). The causal effect of randomization for subject i is the difference in outcomes that would have occurred under treatment and control: $Y_i(1, D_i(1)) - Y_i(0, D_i(0))$. Under some assumptions — the stable unit treatment

value assumption (no interference between units), and random treatment assignment — the causal effect in the trial sample is

$$\frac{\sum_{i=1}^{n}\left[Y_i(1, D_i(1)) - Y_i(0, D_i(0))\right]}{n},$$

and the causal effect in the population is

$$\Upsilon_{\text{ITT}} = \mathbb{E}_{\mathcal{R}}\left[\frac{\sum_{i=1}^{n}\left[Y_i(1, D_i(1)) - Y_i(0, D_i(0))\right]}{n}\right],$$

where n is the sample size in the trial and $\mathbb{E}_{\mathcal{R}}$ refers to the sampling of subjects into the trial. We typically think of the expectation operator $\mathbb{E}_{\mathcal{R}}$ as representing simple random sampling from a notional population of subjects who have the disease in question and are able and willing to participate in a trial.(21) In many medical applications, however, it is clear that the notional population is not the same as the *target* population of subjects who have the disease and would be eligible to take the experimental treatment if it were available.

We can observe a subject's outcome only under the treatment to which he is randomized. For example, if $Z_i = 1$, only $Y_i(1, D_i(1))$ is available, and $Y_i(0, D_i(0))$ is an unobserved potential outcome. For convenience, we denote the observed outcome $Y_i = Z_i Y_i(1, D_i(1)) + (1 - Z_i)Y_i(0, D_i(0))$. An unbiased estimate of the causal effect of randomization Υ_{ITT} is

$$\hat{\Upsilon}_{\text{ITT}} = \frac{\sum_i Z_i Y_i}{\sum_i Z_i} - \frac{\sum_i (1 - Z_i) Y_i}{\sum_i (1 - Z_i)}, \tag{4.1}$$

also known as the *intention-to-treat estimate* of the treatment effect. Therefore we can also designate Υ_{ITT} as the *intention-to-treat estimand*. A valid standard error (SE) of $\hat{\Upsilon}_{\text{ITT}}$ is

$$\text{SE}(\hat{\Upsilon}_{\text{ITT}}) = \sqrt{\frac{s_{Y,0}^2}{\sum_i (1 - Z_i)} + \frac{s_{Y,1}^2}{\sum_i Z_i}},$$

where $s_{Y,0}^2$ and $s_{Y,1}^2$ are the sample variances of Y in the groups randomized to control and experimental treatment, respectively.

4.3.2. The effect of randomization on compliance

The intention-to-treat estimand Υ_{ITT} measures the causal effect of randomization on response, without consideration of compliance. To measure the treatment effect among compliers, it is necessary to consider the effect of treatment assignment on compliance. We denote this estimand as

$$\delta_{\text{ITT}} = \mathbb{E}_{\mathcal{R}}\left[D_i(1) - D_i(0)\right] = \mathbb{E}_{\mathcal{R}}\left[\frac{\sum_{i=1}^{n}\left\{D_i(1) - D_i(0)\right\}}{n}\right].$$

Moreover, define $\pi_1 = \mathbb{E}_{\mathcal{R}}\left[\frac{\sum_i Z_i D_i}{\sum_i Z_i}\right]$ to be the fraction of patients in the treatment group taking experimental treatment, and define $\pi_0 = \mathbb{E}_{\mathcal{R}}\left[\frac{\sum_i (1-Z_i) D_i}{\sum_i (1-Z_i)}\right]$ to be the fraction of patients taking experimental treatment in the control group. The causal effect of randomization on treatment received is then $\delta_{\text{ITT}} = \pi_1 - \pi_0$. Assuming *no defiers* (i.e., $D_i(1) \geq D_i(0)$), this quantity measures the proportion of compliers.

Letting $\hat{\pi}_1$ and $\hat{\pi}_0$ be the estimates of π_1 and π_0, respectively, from trial data, define

$$\hat{\delta}_{\text{ITT}} = \hat{\pi}_1 - \hat{\pi}_0 = \frac{\sum_i Z_i D_i}{\sum_i Z_i} - \frac{\sum_i (1 - Z_i) D_i}{\sum_i (1 - Z_i)},$$

as the ITT estimate of the effect of treatment on adherence. Its standard error is

$$\text{SE}(\hat{\delta}_{\text{ITT}}) = \sqrt{\frac{s_{D,0}^2}{\sum_i (1 - Z_i)} + \frac{s_{D,1}^2}{\sum_i Z_i}},$$

where $s_{D,0}^2$, $s_{D,1}^2$ are the sample variances of D in the control and experimental arms, respectively. As D_i is binary, $s_{D,0}^2 = \hat{\pi}_0(1 - \hat{\pi}_0)$, and $s_{D,1}^2 = \hat{\pi}_1(1 - \hat{\pi}_1)$. When the control arm has no access to the experimental treatment (as in the LRC-CPPT example), $D_i(0) = 0\ \forall i$, and therefore $\hat{\pi}_0 = 0$ and $s_{D,0}^2 = 0$.

4.3.3. The complier average causal effect

To measure the treatment effect among the latent subgroup of compliers, or subjects for whom $D_i(Z_i) = Z_i$, we define the *complier average causal effect (CACE)*

$$\Upsilon_{\text{CACE}} = \mathbb{E}_{\mathcal{R}}\left[Y_i(1) - Y_i(0) | D_i(1) - D_i(0) = 1\right].$$

Under two further assumptions — the exclusion restriction (Y depends on Z only through D), and a nonzero average causal effect on compliance (3, 21, 47) — CACE simplifies to

$$\Upsilon_{\text{CACE}} = \frac{\mathbb{E}_{\mathcal{R}}\left[Y_i(1, D_i(1)) - Y_i(0, D_i(0))\right]}{\mathbb{E}_{\mathcal{R}}\left[D_i(1) - D_i(0)\right]} = \frac{\Upsilon_{\text{ITT}}}{\delta_{\text{ITT}}}.$$

A simple estimate of CACE, known as the *instrumental variables* estimate, is

$$\hat{\Upsilon}_{\text{CACE}} = \frac{\hat{\Upsilon}_{\text{ITT}}}{\hat{\delta}_{\text{ITT}}},$$

with delta method variance

$$\frac{1}{\hat{\delta}^2_{\text{ITT}}}\mathbb{V}(\hat{\Upsilon}_{\text{ITT}}) + \frac{\hat{\Upsilon}^2_{\text{ITT}}}{\hat{\delta}^4_{\text{ITT}}}\mathbb{V}(\hat{\delta}_{\text{ITT}}) - 2\frac{\hat{\Upsilon}_{\text{ITT}}}{\hat{\delta}^3_{\text{ITT}}}\mathbb{C}(\hat{\Upsilon}_{\text{ITT}}, \hat{\delta}_{\text{ITT}}), \tag{4.2}$$

where $\mathbb{V}(\cdot)$ and $\mathbb{C}(\cdot, \cdot)$ are sample variance and covariance operators.(21)

4.4. Interpolation of estimands and estimates

We here discuss methods for projecting causal estimates from a clinical trial population to a target population that differs from it with respect to the distribution of important covariates.

4.4.1. Interpolation with post-stratification

Post-stratification is a traditional projection method for that reweights treatment effects within subsets of the trial population by weights that reflect the proportion of the subset in the target population. Assume that we have recoded all relevant covariates X into a single stratification variable S that takes values s_j, $j = 1, \ldots, J$. As it is necessary to have enough data to estimate the treatment effect separately within each stratum, the presence of continuous or finely measured covariates can complicate analysis. We circumvent the problem of small strata by grouping together similar strata so that all have a workable minimum size.

4.4.1.1. Post-stratification of the ITT estimand

Let $f^{\mathcal{T}}(s)$, $f^{\mathcal{P}}(s)$ represent the distribution of the stratification factor in the trial and target populations, respectively. The ITT estimand in the trial population is $\Upsilon^{\mathcal{T}}_{\text{ITT}} = \sum_j f^{\mathcal{T}}(s_j)\Upsilon_{\text{ITT},j}$, where $\Upsilon_{\text{ITT},j}$ is the estimand in stratum s_j. Similarly, the ITT estimand for the target population is $\Upsilon^{\mathcal{P}}_{\text{ITT}} = \sum_j f^{\mathcal{P}}(s_j)\Upsilon_{\text{ITT},j}$.

Letting $\hat{\Upsilon}_{\text{ITT},j}$ be the ITT estimate from stratum j in the sample, then the clinical trial ITT estimate from Equation (4.1) is approximately

$$\hat{\Upsilon}_{\text{ITT}} \approx \hat{\Upsilon}^{\mathcal{T}}_{\text{ITT}} = \sum_{j=1}^{J} f^{\mathcal{T}}(s_j)\hat{\Upsilon}_{\text{ITT},j}.$$

The two estimates will may differ slightly due to small imbalances of randomization within strata. An interpolated estimate for the ITT estimand in the target population is then(29)

$$\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}} = \sum_{j=1}^{J} f^{\mathcal{P}}(s_j)\hat{\Upsilon}_{\text{ITT},j}. \tag{4.3}$$

In order for there to be a difference in estimands ($\mathbb{E}[\hat{\Upsilon}^{\mathcal{T}}_{\text{ITT}}] \neq \mathbb{E}[\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}}]$), there must be an interaction between treatment and stratum, and the trial and target distributions of S must differ.(29)

4.4.1.2. Post-stratification estimate of CACE

Similarly, the causal effect of randomization on treatment received in the trial sample is $\delta^{\mathcal{T}}_{\text{ITT}} = \mathbb{E}_{\mathcal{R}}\left[\sum_j f^{\mathcal{T}}(s_j)\delta_{\text{ITT},j}\right]$, which we compare to the causal effect in the target population $\delta^{\mathcal{P}}_{\text{ITT}} = \mathbb{E}_{\mathcal{R}}\left[\sum_j f^{\mathcal{P}}(s_j)\delta_{\text{ITT},j}\right]$. Letting $\hat{\delta}_{\text{ITT},j}$ be the estimated effect on compliance in stratum j, we obtain the post-stratification projection to the target population

$$\hat{\delta}^{\mathcal{P}}_{\text{ITT}} = \sum_{j=1}^{J} f^{\mathcal{P}}(s_j)\hat{\delta}_{\text{ITT},j}. \tag{4.4}$$

Under the assumptions of IV estimation, for the target population we get $\Upsilon^{\mathcal{P}}_{\text{ITT}} = \delta^{\mathcal{P}}_{\text{ITT}} \times \Upsilon^{\mathcal{P}}_{\text{CACE}}$. Therefore the IV estimand of CACE in the target population is the ratio

$$\Upsilon^{\mathcal{P}}_{\text{CACE}} = \frac{\Upsilon^{\mathcal{P}}_{\text{ITT}}}{\delta^{\mathcal{P}}_{\text{ITT}}} = \frac{\sum_{j=1}^{J} f^{\mathcal{P}}(s_j)\Upsilon_{\text{ITT},j}}{\sum_{j=1}^{J} f^{\mathcal{P}}(s_j)\delta_{\text{ITT},j}}$$

To estimate CACE in the target population, one must therefore interpolate the numerator and denominator separately, leading to

$$\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}} = \frac{\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}}}{\hat{\delta}^{\mathcal{P}}_{\text{ITT}}}. \tag{4.5}$$

4.4.1.3. Standard errors of the interpolated estimates

By independence across strata, standard errors of $\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}}$ and $\hat{\delta}^{\mathcal{P}}_{\text{ITT}}$ are

$$\mathrm{SE}(\hat{\Upsilon}^{\mathcal{P}}_{\mathrm{ITT}}) = \sqrt{\sum_{j=1}^{J} [f^{\mathcal{P}}(s_j)]^2 \,\mathrm{SE}^2\left(\hat{\Upsilon}_{\mathrm{ITT},j}\right)}, \tag{4.6}$$

$$\mathrm{SE}(\hat{\delta}^{\mathcal{P}}_{\mathrm{ITT}}) = \sqrt{\sum_{j=1}^{J} [f^{\mathcal{P}}(s_j)]^2 \,\mathrm{SE}^2\left(\hat{\delta}_{\mathrm{ITT},j}\right)}. \tag{4.7}$$

The delta-method standard error in Equation (4.2) can be unstable. A practical alternative, which we have implemented in our examples, is to compute the SE by the bootstrap, re-sampling triples (Z_i, D_i, Y_i) within strata. To make the results reproducible, we recommend resampling at least 5,000 times.

4.4.2. Interpolation with inverse inclusion probability weighting

Stuart *et al.* (48) proposed an interpolation method that mimics propensity score analysis. First one computes, for each subject in the trial, the probability of being selected into the trial; we denote the inverse of this probability as W_i^{IIPW}, which we then use aa an analysis weight. The *inverse inclusion probability weight (IIPW)* estimate of the ITT effect of Z on Y is then

$$\hat{\Upsilon}^{\mathcal{P},\mathrm{IIPW}}_{\mathrm{ITT}} = \frac{\sum_i Z_i Y_i W_i^{\mathrm{IIPW}}}{\sum_i Z_i W_i^{\mathrm{IIPW}}} - \frac{\sum_i (1-Z_i) Y_i W_i^{\mathrm{IIPW}}}{\sum_i (1-Z_i) W_i^{\mathrm{IIPW}}}. \tag{4.8}$$

We have shown that this estimate is approximately equal to the post-stratified estimate when using the same set of discrete stratifiers. The two approaches give identical results when we have equal randomization in all strata.(29)

Similarly, the IIPW estimate of the causal effect of treatment on adherence is

$$\hat{\delta}^{\mathcal{P},\mathrm{IIPW}}_{\mathrm{ITT}} = \frac{\sum_i Z_i D_i W_i^{\mathrm{IIPW}}}{\sum_i Z_i W_i^{\mathrm{IIPW}}} - \frac{\sum_i (1-Z_i) D_i W_i^{\mathrm{IIPW}}}{\sum_i (1-Z_i) W_i^{\mathrm{IIPW}}}.$$

Thus, the IIPW estimate of CACE is the ratio of the two interpolated causal estimates:

$$\hat{\Upsilon}^{\mathcal{P},\text{IIPW}}_{\text{CACE}} = \frac{\hat{\Upsilon}^{\mathcal{P},\text{IIPW}}_{\text{ITT}}}{\hat{\delta}^{\mathcal{P},\text{IIPW}}_{\text{ITT}}}.$$

We will discuss the SE of interpolated estimates from IIPW in Section 4.4.4.

4.4.3. Interpolation with inclusion probability stratification

One can also use the trial inclusion probabilities to conduct a stratified analysis.(38) With this approach, one creates strata based on the estimated inclusion scores, then conducts a post-stratified analysis using these newly derived strata. An advantage of this method is that, like propensity score analysis for observational data, it appears to work well with as few as five created strata.(42) It also avoids issues the potential instability of the weighted analysis, and the need to artificially combine strata to avoid small numbers.

Formally, we define an artificial stratification variable S^{IPS} that takes values $s^{\text{IPS}}_k, k = 1, \ldots K$, based on quantiles of the inclusion probabilities. Then we can compute an estimated ITT effect $\hat{\Upsilon}_{\text{ITT},k}$ in each s^{IPS}_k from the trial data, and apply Equations (4.3) based on the distribution of S^{IPS} in the target population, $f^{\mathcal{P}}(s^{\text{IPS}})$:

$$\hat{\Upsilon}^{\mathcal{P},\text{IPS}}_{\text{ITT}} = \sum_{k=1}^{K} f^{\mathcal{P}}(s^{\text{IPS}}_k)\hat{\Upsilon}_{\text{ITT,k}}.$$

If one uses the quantiles of inclusion probabilities as observed in the trial, the number of observations is balanced in each stratum. Alternatively, if one defines quantiles from the target populations, we have $f^{\mathcal{P}}(s^{\text{IPS}}_k) = 1/K \;\forall\; k$.

One can readily extend this approach to estimate the CACE: First interpolate within-stratum estimates of compliance $\hat{\delta}_{\text{ITT},k}$ to get $\hat{\delta}^{\mathcal{P},\text{IPS}}_{\text{ITT}}$; then use Equation (4.5) to obtain the

interpolated CACE:

$$\hat{\Upsilon}^{\mathcal{P},\mathrm{IPS}}_{\mathrm{CACE}} = \frac{\hat{\Upsilon}^{\mathcal{P},\mathrm{IPS}}_{\mathrm{ITT}}}{\hat{\delta}^{\mathcal{P},\mathrm{IPS}}_{\mathrm{ITT}}}.$$

We will discuss SE estimation in Section 4.4.4.

4.4.4. Implementation

4.4.4.1. Selection of stratifiers

The first step is to identify a relevant set of covariates. As a rule, one should include in the stratification any covariate that is associated with heterogeneity of randomization effects on either the outcome or treatment received.

To check for heterogeneity of treatment effect by covariates, one can use a model including randomization, covariates (X), and interactions between randomization and covariates on response: (29)

$$Y = \alpha^{\dagger} + X\beta^{\dagger} + Z\Upsilon^{\dagger} + ZX\gamma^{\dagger} + E, \tag{4.9}$$

where E is Normally distributed with variance σ^2_y. Note that the treatment effect $\Upsilon^{\dagger}$ is not the same as Υ_{ITT}, which is the treatment effect averaged over the covariate distribution.

When control subjects have access to the experimental treatment (as in the New York School Choice Study), we can similarly conduct a regression of D on X, Z, and the interaction of X and Z to identify covariates for which treatment effects are heterogeneous:

$$D = \alpha^{\dagger}_d + X\beta^{\dagger}_d + Z\Upsilon^{\dagger}_d + ZX\gamma^{\dagger}_d + E_d. \tag{4.10}$$

Although one would usually use a generalized linear model with a logit link function in this context, because the treatment effects on adherence are to be measured on the probability

scale, it is preferable to use an identity link for this model. (1) Moreover, we use the estimated residual variance from this model, call it $\hat{\sigma}_d^2$, in subsequent SE calculations.

When subjects in the control group have no access to the experimental treatment (as in LRC-CPPT), any covariate that is associated with D in the experimental arm has an *a fortiori* interaction with randomization. Therefore in this case we evaluate covariate effects from a reduced model that excludes interactions, using data on the experimental arm only.

4.4.4.2. Standard errors for the IIPW estimates

We have observed that one can compute the IIPW estimate of the effect of randomization on Y *via* a weighted regression of Y on Z using the inverse inclusion probabilities as the weights. Unfortunately, the default SE from this approach may be many overestimate variability, as it does not account for heterogeneity of effects across values of X. Instead, one can rewrite the estimate as a linear function of the Y_i values, $\hat{\Upsilon}_{\text{ITT}}^{\mathcal{P},\text{IIPW}} = \sum_i A_i Y_i$ with appropriately chosen A_i values. An efficient SE is then $\text{SE}(\hat{\Upsilon}_{\text{ITT}}^{\mathcal{P},\text{IIPW}}) = \hat{\sigma}_y \sqrt{\sum_i A_i^2}$, where $\hat{\sigma}_y^2$ is the estimated variance from Equation (4.9). This renders the SE from this approach roughly equal to the SE from post-stratification with an equivalent set of included covariates. To compute the SE for $\hat{\delta}_{\text{ITT}}^{\mathcal{P},\text{IIPW}}$ one uses a similar approach, but substituting $\hat{\sigma}_d$ from the linear regression of D in the model for heterogeneity.(1) Because the delta method variance for the ratio of treatment effects can be inaccurate and unstable, in particular if the denominator effect is modest, it is preferable to compute SEs by the bootstrap.

4.4.4.3. Standard errors for the inverse probability stratification estimates

To compute the SE of $\hat{\Upsilon}_{\text{ITT}}^{\mathcal{P},\text{IPS}}$, one simply computes the SE using the post-stratification approach, but substituting $\hat{\sigma}_y$ from Model (4.9) as the standard deviation term.(29) Similarly, to estimate the SE of $\hat{\delta}_{\text{ITT}}^{\mathcal{P},\text{IPS}}$ one substitutes $\hat{\sigma}_d$ from Model (4.10) as the standard deviation

term. In computing the SE of $\hat{\Upsilon}^{\mathcal{P},\mathrm{IPS}}_{\mathrm{CACE}}$ it is generally preferable to use the bootstrap, for reasons described above.

4.5. Simulation study

We conducted a simulation study to i) compare interpolated causal estimates from post-stratification and two approaches using inclusion probabilities; and ii) demonstrate how selection of covariates for post-stratification affects the results. We created a synthetic target population by making six copies of the LRC-CPPT trial dataset (randomization, compliance, and covariates only).

To generate a variable for actual treatment received, we first created a continuous variable C so that compliance depend on randomization, smoking status, and age:

$$\mathbb{E}(C) = 1.8 \times Z - 0.05 \times \texttt{smoke} + 0.05 \times \texttt{age}.$$

In the trial, we assumed patients assigned to take placebo did not have access to the treatment, so we did not include interaction terms between randomization and covariates on compliance. We added independent $\mathbb{N}(0, 0.36^2)$ deviates to C and set $D = 1$ if $C > 4.1$, so that all the patients in the control group have $D = 0$, and the fraction of compliers was close to 52% in the synthetic target population, as observed in the actual clinical trial dataset.

Next we generated an outcome for which education, smoking, and age exhibit no, moderate, and strong interaction, respectively, with actual treatment received:

$$\begin{aligned} Y &= -12.2 + 11 \times D - 2 \times \texttt{smoke} + 0.3 \times \texttt{age} + 0.5 \times \texttt{education} \\ &\quad -0.2 \times Z \times \texttt{smoke} + 0.5 \times Z \times \texttt{age} + E, \end{aligned}$$

where E represented independent $\mathbb{N}(0, 3^2)$ deviates. We divided `education` into three groups — less than high school, high school graduate, and some college and above. We randomly selected participants for the synthetic trial population with the logistic model

$$\text{logit}(p) = -4.59 - 0.5 \times \texttt{smoke} + 0.05 \times \texttt{age} + 0.6 \times \texttt{education};$$

where p is the trial inclusion probability. From the resulting synthetic population of 22,704 subjects we sampled 3,792 to become the synthetic trial dataset. We used continuous age for trial inclusion score analysis, but we divided age into three categories for post-stratification: < 44, 44–50, and ≥ 51.

Table 4.1 shows the estimated causal effects from the synthetic trial, the synthetic target population, inverse inclusion probability weighting with true and estimated trial inclusion probabilities, inclusion probability stratification, and post-stratification with all possible combinations of education, smoking and age group. The ground truth in the synthetic population is $\hat{\delta}^{\mathcal{P}}_{\text{ITT}} = 0.52$, $\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}} = 29.02$, and $\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}} = 56.11$. These estimates differed from the results using the synthetic trial data directly. $\hat{\delta}^{\mathcal{P}}_{\text{ITT}}$ from all approaches using inclusion probabilities were close to the ground truth. Interpolated estimates using both the true trial inclusion scores and estimated trial inclusion scores were within 1 SE of the ground truth.

Post-stratification including age group, the covariate with the strongest interaction with treatment, returned causal estimates that were closest to the ground truth. Yet even these estimates were off by 2 to 3 SE. Presumably this is a consequence of having to arbitrarily group the age variable to render it categorical, as the inclusion probability methods did not exhibit this bias. Besides, the simulated treatment effects show a strong interaction with age group with relatively small SEs, discretization of age causes further issues.

4.6. Real-data applications

4.6.1. LRC-CPPT

We conducted stepwise variable selection in two models: A regression of cholesterol drop (Y) on randomization, covariates, and their interaction in the entire data set; and a regression of treatment received (D) on randomization and covariates in the subgroup randomized to cholestyramine ($Z = 1$). Both analyses identified age and smoking status as important covariates; therefore we use these variables in all subsequent interpolations. We also included race, whose interaction with treatment for Y is large but not statistically significant. In all analyses, we defined compliance to mean taking at least 80% of prescribed packets, leading to 52% compliance in the cholestyramine arm.

Estimated causal effects by smoking status appear in Table 4.2. The proportion of smokers in LRC-CPPT was 76%, whereas NHANES estimated that 45% of American males with hypercholesterolemia are ever-smokers.(8) Applying Equations (4.3), (4.4), and (4.5) with the distribution of smokers from NHANES we get $\hat{\delta}^{\mathcal{P}}_{\text{ITT}} = 0.53$ (SE=0.01), $\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}} = 28.02$ (0.90), $\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}} = 52.55$ (1.42), compared to $\hat{\delta}^{\mathcal{T}}_{\text{ITT}} = 0.52$ (0.01), $\hat{\Upsilon}^{\mathcal{T}}_{\text{ITT}} = 26.91$ (0.80), $\hat{\Upsilon}^{\mathcal{T}}_{\text{CACE}} = 51.53$ (1.47) in the trial.

We took as our target population males aged 34–60 in NHANES who met current standards for hypercholesterolemia (total cholesterol $\geq$ 200 mg/dL; LDL cholesterol $\geq$ 100 mg/dL; triglycerides $\leq$ 300 mg/dL). This led to a population of 2,118 subjects. To compute trial inclusion probabilities, we weighted the subjects in the target population with their NHANES sampling weights, which reflect representation of these men in the US population. We assigned LRC-CPPT participants each a weight of 1, as they represent only themselves. We then estimated a weighted logistic model with age and smoking status, taking $T = 1$ as the outcome for trial participation for the LRC-CPPT members and $T = 0$ as the outcome

for the NHANES men. The inclusion probabilities were the predicted probabilities of $T = 1$ for the trial participants.

For the post-stratification, we grouped subjects by age quantiles so that the number of patients in each category would be similar. We conducted analyses with $2, 5$, and 10 subgroups on age to see the effect of the number of subgroups on interpolated estimates. The total numbers of strata, after combination with smoking status, are therefore $J = 4$, 10, and 20, respectively. Table 4.3 displays projected estimates of treatment effects, post-stratified by age group and smoking status. The number of strata had little effect on the estimates for $\hat{\delta}^{\mathcal{P}}_{\text{ITT}}$ and $\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}}$, but the SE for $\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}}$ increased with J.

4.6.2. NY School Choice Experiment

A stepwise variable selection for reading score identified no variable with significant interaction with randomization, whereas both race and sex had significant interactions with randomization for math score. A stepwise variable selection on compliance identified race as having a significant interaction with treatment. Therefore in subsequent analyses we project based on race and sex. In analysis limited to two binary covariates, the three interpolation methods will give identical results, so we henceforth present only post-stratification estimates. We base our interpolations on 2000 US Census data for New York City and the entire United States.(11, 36)

Table 4.4 shows that males have higher improvement on reading, whereas females and blacks have higher improvement in math. The proportion of blacks in the study (49.0%) far exceeds that in the New York City population, which was 29.9% in the 2000 census.(36)

Interpolating to the New York race distribution, the resulting $\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}}$ is 3.88 (SE=2.54) for reading and 2.95 (SE=2.83) for math (Table 4.5). Because the causal estimates differ by race, the interpolated estimate for math deviates from the trial estimate. Nevertheless the

difference is within 1 SE.

When interpolating math score for race and sex simultaneously, the overall CACE effect for non-blacks is close to 0, as the strong positive effect for females and strong negative effect for males cancel out. The 2000 census in NYC showed that for children aged 6–9 in New York City, 50.7% of non-Hispanic blacks were males, and 51.1% of non-blacks were males.(36) Using these proportions, the values of $\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}}$ are 4.07 (SE=2.49) for reading and 2.35 (SE=2.84) for math. The math estimates differ from the overall estimates using the trial data, but the difference is less for reading scores. As we identified strong interactions between randomization and race and sex for math scores, the interpolated math scores show further deviation from the trial estimates when we post-stratify by both race and sex. When projected to the entire US population based on the 2000 Census, the causal estimates for math further deviated from the trial, suggesting that applying the treatment across the country would lead to modest improvements in performance. Gains for reading would be sustained, however.

4.7. Discussion

We have described projection of causal estimates from a randomized trial to a target population using post-stratification, inverse inclusion probability weighting, and inclusion probability stratification. All the methods require that subjects in the trial are similar to those in the target population, other than on a set of measured covariates, and that information is available on these covariates in the population. Furthermore, we assume that compliance behavior for patients in the trial is similar to those in the target population under the same subgroup, so that the estimated fraction of compliers in the trial can directly apply to the target population in the same subgroup.

Before interpolation, one should use the trial data to identify covariates that display a strong interaction with randomization on either the response or compliance. One can also

include variables that are likely to evince heterogeneity but may not demonstrate statistical significance due to sample size limitations. In our simulation study, the estimates using methods involving inclusion scores were close to the ground truth. Post-stratification results was less effective, presumably due to its reliance on categorical variables only. Although post-stratification is less flexible, we can use it to conduct a sensitivity analysis with the joint distribution of confounders in the target population when subject-level population data are unavailable.

We have shown that the weights implied in post-stratification is similar to the weights from inverse of trial inclusion scores using the same set of discrete stratifiers. Therefore, if all variables are discrete, the three methods will give similar answers.

If subject-level data are not available in the target population, one can create a pseudo-data set consisting of individuals weighted by their proportions in the population and use this as the basis of a model for computing trial inclusion probabilities. Again, if all predictors are categorical, this analysis should closely resemble the post-stratification analysis.

In both data examples, we defined the response variable as the difference between follow-up and baseline measurements.(5, 14) This is less efficient compared with ANCOVA including baseline measure, which is more preferable.(15, 45)

Control-group patients in a clinical trial may still observe a positive effect due to benefits of trial participation such as reduced health care cost and frequent evaluations.(52) Nevertheless, there is evidence showing that treatment effects from randomized experiments are similar to observational studies, (37, 40) which justifies the use of projection methods.

An important general issue is that failure to adjust for unmeasured confounders — variables that are both associated with participation and have interactions with treatment — can lead to biased projections. In the New York School Choice Experiment example, potential confounders include variables such as school transportation, the availability and quality of after-school programs, etc., that are likely to differ across the country and affect test out-

comes. These variables are not available in the dataset or the Census. An important future direction is to address this concern *via* sensitivity analysis.(35)

4.8. Acknowledgements

We thank John Barnard and Donald Rubin for providing the New York School Choice dataset. We obtained the LRC-CPPT data from the NHLBI Biologic Specimen & Data Repository Information Coordinating Center. Our work does not necessarily reflect the opinions or views of the LRC-CPPT or the NHLBI.

Table 4.1. Interpolated causal estimates (SE) from simulated data based on LRC-CPPT; outcome is actual treatment received for $\hat{\delta}_{\text{ITT}}$, and drop in cholesterol (mg/dL) for $\hat{\Upsilon}_{\text{ITT}}$ and $\hat{\Upsilon}_{\text{CACE}}$. Selection into the trial is based on age, smoking status, and education. Compliance is based on age and smoking status.

Method	$\hat{\delta}^{\mathcal{T}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{T}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{T}}_{\text{CACE}}$
Synthetic trial (N = 3,792)	0.59 (0.01)	30.54 (0.24)	51.81 (0.70)
	$\hat{\delta}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}}$
Synthetic population (N = 22,704)	0.52	29.02	56.11
Trial inclusion score analysis	$\hat{\delta}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}}$
IIPW with true score	0.52 (0.01)	29.03 (0.25)	55.67 (1.20)
IIPW with true score, corrected variance	0.52 (0.01)	29.03 (0.20)	55.67 (1.20)
IIPW with estimated score	0.53 (0.01)	29.16 (0.25)	55.31 (1.15)
IIPW with estimated score, corrected variance	0.53 (0.01)	29.16 (0.15)	55.31 (1.15)
IPS analysis	0.52 (0.01)	28.94 (0.25)	55.71 (1.25)
IPS analysis, correct variance	0.52 (0.01)	28.94 (0.16)	55.71 (1.25)
Post-stratification	$\hat{\delta}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}}$
Education	0.60 (0.01)	30.82 (0.25)	51.24 (0.67)
Smoking	0.59 (0.01)	30.48 (0.24)	52.02 (0.70)
Age group	0.54 (0.01)	29.48 (0.16)	54.72 (0.80)
Education+smoking	0.60 (0.01)	30.76 (0.25)	51.42 (0.68)
Education+age group	0.54 (0.01)	29.50 (0.17)	54.66 (0.80)
Smoking+age group	0.54 (0.01)	29.48 (0.16)	54.84 (0.80)
Education+smoking+age group	0.54 (0.01)	29.54 (0.17)	54.58 (0.79)

Table 4.2. Causal effect estimates (SE) by smoking status from LRC-CPPT; outcome is the treatment received for $\hat{\delta}_{\text{ITT}}$, and drop in cholesterol (mg/dL) for $\hat{\Upsilon}_{\text{ITT}}$ and $\hat{\Upsilon}_{\text{CACE}}$.

	Smoking	
	Non-smoker	Smoker
	(N = 904)	(N = 2880)
$\hat{\delta}_{\text{ITT},j}$	0.56 (0.02)	0.51 (0.01)
$\hat{\Upsilon}_{\text{ITT},j}$	30.96 (1.66)	25.62 (0.91)
$\hat{\Upsilon}_{\text{CACE},j}$	55.06(2.68)	50.29 (1.74)

Table 4.3. Interpolated causal effect (SE) based on the distribution of smoking status and age in NHANES. Outcome is treatment received for $\hat{\delta}_{\text{ITT}}$ and drop in cholesterol (mg/dL) for $\hat{\Upsilon}_{\text{ITT}}$ and $\hat{\Upsilon}_{\text{CACE}}$.

	$\hat{\delta}^{\mathcal{T}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{T}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{T}}_{\text{CACE}}$
Trial	0.52 (0.01)	26.91 (0.80)	51.53 (1.47)
Analysis	$\hat{\delta}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}}$
Post-stratification, $J = 4$	0.53 (0.01)	28.33 (0.91)	53.14 (1.43)
Post-stratification, $J = 10$	0.53 (0.01)	28.27 (0.93)	53.16 (2.19)
Post-stratification, $J = 20$	0.53 (0.01)	28.29 (0.94)	53.08 (2.82)
IIPW	0.54 (0.01)	28.15 (0.79)	52.57 (1.42)
IPS	0.53 (0.01)	28.01 (0.87)	52.46 (1.43)

Table 4.4. Causal effect estimates (SE) by stratum from the New York School Choice Experiment; the outcome is actual attendance of private school for $\hat{\delta}_{\text{ITT}}$, and change in grade-normed national centile ranking for $\hat{\Upsilon}_{\text{ITT}}$ and $\hat{\Upsilon}_{\text{CACE}}$.

	Male ($n = 349$)		Female ($n = 366$)	
	Reading	Math	Reading	Math
$\hat{\hat{\delta}}_{\text{ITT},j}$	0.75 (0.03)		0.73 (0.04)	
$\hat{\Upsilon}_{\text{ITT},j}$	4.81 (1.98)	0.46 (2.21)	0.84 (1.81)	5.66 (2.12)
$\hat{\Upsilon}_{\text{CACE},j}$	6.42 (2.64)	0.61 (2.95)	1.16 (2.48)	7.78 (2.94)

	Non-black ($n = 365$)		Black ($n = 350$)	
	Reading	Math	Reading	Math
$\hat{\hat{\delta}}_{\text{ITT},j}$	0.67 (0.04)		0.81 (0.03)	
$\hat{\Upsilon}_{\text{ITT},j}$	2.22 (1.90)	0.35 (2.27)	4.01 (1.85)	6.20 (2.03)
$\hat{\Upsilon}_{\text{CACE},j}$	3.32 (2.82)	0.52 (3.40)	4.96 (2.30)	7.67 (2.54)

Table 4.5. Interpolated results (SE) from the New York School Choice Experiment; the outcome is actual attendance of private school for $\hat{\delta}_{\text{ITT}}$, and change in grade-normed national centile ranking for $\hat{\Upsilon}_{\text{ITT}}$ and $\hat{\Upsilon}_{\text{CACE}}$.

		Reading		
Stratifier	Distribution	$\hat{\delta}^{\mathcal{T}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{T}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{T}}_{\text{CACE}}$
—	Trial	0.74 (0.02)	2.80 (1.34)	3.79 (1.80)
		$\hat{\delta}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}}$
Sex	NYC 2000 Census	0.74 (0.03)	2.87 (1.34)	3.88 (2.56)
	US 2000 Census	0.74 (0.03)	2.88 (1.34)	3.89 (2.55)
Race	NYC 2000 Census	0.71 (0.03)	2.76 (1.44)	3.88 (2.54)
	US 2000 Census	0.69 (0.03)	2.50 (1.63)	3.62 (2.56)
Race & sex	NYC 2000 Census	0.71 (0.03)	2.89 (1.44)	4.07 (2.49)
	US 2000 Census	0.69 (0.03)	2.70 (1.64)	3.89 (2.48)

		Math		
Stratifier	Distribution	$\hat{\delta}^{\mathcal{T}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{T}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{T}}_{\text{CACE}}$
—	Trial	0.74 (0.02)	3.07 (1.53)	4.15 (2.07)
		$\hat{\delta}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{ITT}}$	$\hat{\Upsilon}^{\mathcal{P}}_{\text{CACE}}$
Sex	NYC 2000 Census	0.74 (0.03)	3.01 (1.53)	4.07 (2.87)
	US 2000 Census	0.74 (0.03)	3.01 (1.53)	4.07 (2.81)
Race	NYC 2000 Census	0.71 (0.03)	2.10 (1.70)	2.95 (2.83)
	US 2000 Census	0.69 (0.03)	1.26 (1.94)	1.82 (2.85)
Race & sex	NYC 2000 Census	0.71 (0.03)	1.67 (1.72)	2.35 (2.84)
	US 2000 Census	0.69 (0.03)	0.73 (1.96)	1.05 (2.84)

CHAPTER 5

CONCLUSION

In this dissertation, we have proposed a rule-based phase I dose-finding design for oncology studies. In addition, we have projected the treatment effects from a randomized clinical trial to a target population. One innovation aspect of this project is interpolation in medical trials considering compliance.

In Chapter 2, the proposed design is flexible based on the target toxicity probability; it is more efficient on dose assignment so that more patients can be treated near the potential maximum tolerated dose, which will be recommended for future trials; and the design is simple to use. We have developed a package in `R` to make it more accessible.

The design is targeted for phase I oncology trials in which the toxicity increases with increasing doses. The aim is to obtain preliminary information on safety fast with few number of patients. In cases where treatments do not have monotonic dose toxicity relationship, other designs are better suited.

Rule-based designs use information from patients treated at the maximum tolerated dose only, whereas model-based designs uses information from all patients to identify a dose that delivers a targeted dose limiting toxicity rate. This is more accurate and efficient; but it requires a model based on little data to begin with, then conducts analysis after each response to recommend dose for the next patient. It requires more time from statisticians and clinicians. Thus, there is still a continuing role for rule-based designs, especially in institutions where statistical resources are limited.

After getting preliminary safety and efficacy information from early-stage trials, more patients are randomized in late-stage trials for estimating the treatment effect. In Chapter 3 and 4, we use results from late-stage clinical trials to project to the target population where the same treatment can be applied. When we include stratification factors that show strong interaction on outcome or compliance, the interpolated estimates can differ substantially from the trial estimates. Post-stratification can work with just joint distribution of stratifiers. Approaches using inclusion probabilities are more flexible and accurate on modeling trial inclusion, but it requires individual-level data that is rarely available for medical trials. In addition, these approaches require additional step for getting the standard error of estimates accounting for heterogeneity of causal effect by covariates, which is automatically built-in for post-stratification.

To conduct interpolation, the patients in the trial are similar to those in the target population, other than on a set of measured covariates. A future direction for this research is to combine information from similar trials to include a larger group of subjects for interpolation. One example would be to combine information from other trials on cholestyramine that includes females, so that the target population can include male and female patients.

APPENDIX A

APPENDIX OF CHAPTER 3

A.1. Appendix A. Equivalence of IPW and post-stratification with discrete S

Define $Y_{ij},\ i = 1, \dots n_j,\ j = 1, \dots J$, to be the outcome from subject i in stratum j of the trial sample. Let n_j and N_j represent the total numbers of subjects in stratum j in the trial and target populations, respectively. Thus, $n = \sum_{j=1}^{J} n_j$ is the trial size; $N = \sum_{j=1}^{J} N_j$ is the population size; $f^{\mathcal{T}}(s_j) = n_j/n$; and $f^{\mathcal{P}}(s_j) = N_j/N$. The inverse-probability-weighted treatment effect is

$$\hat{\Upsilon}^{\mathcal{P},\text{IPW}} = \frac{\sum_{i,j} Z_{ij} Y_{ij} W_{ij}^{\text{IPW}}}{\sum_{i,j} Z_{ij} W_{ij}^{\text{IPW}}} - \frac{\sum_{i,j} (1 - Z_{ij}) Y_{ij} W_{ij}^{\text{IPW}}}{\sum_{i,j} (1 - Z_{ij}) W_{ij}^{\text{IPW}}},$$

where the inverse trial inclusion probability is

$$W_{ij}^{\text{IPW}} = \frac{N_j}{n_j}.$$

Moreover, $\sum_{i,j} W_{ij}^{\text{IPW}} = \sum_j \frac{n_j N_j}{n_j} = \sum_j N_j = N$.

Let n_{j1} and n_{j0} be the number of patients in stratum j in the experimental and control arms, respectively. Note that the inverse-inclusion-probability-weighted estimate is a linear function of the observations: $\hat{\Upsilon}^{\mathcal{P},\text{IPW}} = \sum_{i,j} A_{ij} Y_{ij}$, where,

$$A_{ij} = \begin{cases} \frac{N_j/n_j}{\sum_k N_k n_{k1}/n_k}, & \text{if } Z_{ij} = 1, \\ -\frac{N_j/n_j}{\sum_k N_k n_{k0}/n_k}, & \text{if } Z_{ij} = 0. \end{cases}$$

Similarly, the post-stratified estimated $\hat{\Upsilon}^{\mathcal{P}} = \sum_{i,j} B_{ij} Y_{ij}$, where,

$$B_{ij} = \begin{cases} \frac{N_j}{N n_{j1}}, & \text{if } Z_{ij} = 1, \\ -\frac{N_j}{N n_{j0}}, & \text{if } Z_{ij} = 0. \end{cases}$$

Under equal randomization fractions in each stratum, i.e., $n_{j0}/n_j = n_{j1}/n_j = 1/2 \ \forall j$, the methods give identical weights:

$$A_{ij} = \frac{2N_j}{n_j N} = B_{ij}.$$

In practice, the weights in the two approaches will differ slightly due to accidental imbalances in randomization fractions across strata.

BIBLIOGRAPHY

[1] Agresti, A., and Kateri, M. *Categorical data analysis*. Springer, 2011.

[2] Ahn, C. An evaluation of phase i cancer clinical trial designs. *Statistics in medicine 17*, 14 (1998), 1537–1549.

[3] Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association 91*, 434 (1996), 444–455.

[4] Babb, J., Rogatko, A., and Zacks, S. Cancer phase i clinical trials: efficient dose escalation with overdose control. *Statistics in medicine 17*, 10 (1998), 1103–1120.

[5] Barnard, J., Frangakis, C. E., Hill, J. L., and Rubin, D. B. Principal Stratification Approach to Broken Randomized Experiments. *Journal of the American Statistical Association 98*, 462 (2003), 299–323.

[6] Braun, T. M. The current design of oncology phase i clinical trials: progressing from algorithms to statistical models. *Chinese clinical oncology 3*, 1 (2014).

[7] Campbell, D. T. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin 54* (1957), 297–312.

[8] Centers for Disease Control and Prevention (CDC). Nationl health and nutrition examination survey data, 2018.

[9] Cheung, Y. K., and Chappell, R. Sequential designs for phase i clinical trials with late-onset toxicities. *Biometrics 56*, 4 (2000), 1177–1182.

[10] Cole, S. R., and Stuart, E. A. Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology 172*, 1 (2010), 107–115.

[11] Data Access and Dissemination Systems. American factfinder, 2000.

[12] Data Access and Dissemination Systems. American factfinder, 2010.

[13] Edler, L., and Burkholder, I. Overview of phase i trials. In *Handbook of Statistics in Clinical Oncology, Second Edition*. Chapman and Hall/CRC, 2005, pp. 23–50.

[14] Efron, B., and Feldman, D. Compliance as an Explanatory Variable in Clinical Trials. *Journal of the American Statistical Association 86*, 413 (1991), 9–17.

[15] Fleiss, J. L. *The Design and Analysis of Clinical Experiments*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1986.

[16] Frangakis, C. E., and Rubin, D. B. Principal Stratification in Causal Inference. *Biometrics 58*, 1 (2002), 21–29.

[17] Goetghebeur, E., and Molenberghs, G. Causal Inference in a Placebo-Controlled Clinical Trial With Binary Outcome and Ordered Compliance. *Journal of the American Statistical Association 91*, 435 (1996), 928–934.

[18] Hill, J., Rubin, D. B., and Thomas, N. The design of the new york school choice scholarship program evaluation. *Validity and Social Experimentation: Donald Campbell's Legacy 1* (2000), 155–180.

[19] Huang, B., Bycott, P., and Talukder, E. Novel dose-finding designs and considerations on practical implementations in oncology clinical trials. *Journal of biopharmaceutical statistics 27*, 1 (2017), 44–55.

[20] Imbens, G. The role of the propensity score in estimating dose-response functions. *Biometrika 87*, 3 (2000), 706–710.

[21] Imbens, G. W., and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[22] Ivanova, A. Escalation, group and a+ b designs for dose-finding trials. *Statistics in Medicine 25*, 21 (2006), 3668–3678.

[23] Ji, Y., Li, Y., and Nebiyou Bekele, B. Dose-finding in phase i clinical trials based on toxicity probability intervals. *Clinical Trials 4*, 3 (2007), 235–244.

[24] Ji, Y., Liu, P., Li, Y., and Nebiyou Bekele, B. A modified toxicity probability interval method for dose-finding trials. *Clinical Trials 7*, 6 (2010), 653–663.

[25] Ji, Y., and Wang, S.-J. Modified toxicity probability interval design: a safer and more reliable method than the 3+ 3 design for practical phase i trials. *Journal of Clinical Oncology 31*, 14 (2013), 1785.

[26] Jin, H., and Rubin, D. B. Principal Stratification for Causal Inference With Extended Partial Compliance. *Journal of the American Statistical Association 103*, 481 (2008), 101–111.

[27] Julian, D., Jackson, F., Prescott, R., and Szekely, P. Controlled trial of sotalol for one year after myocardial infarction. *The Lancet 319* (1982), 1142–1147.

[28] Le Tourneau, C., Lee, J. J., and Siu, L. L. Dose escalation methods in phase i cancer clinical trials. *JNCI: Journal of the National Cancer Institute 101*, 10 (2009), 708–720.

[29] Li, S., and Heitjan, D. F. Projecting clinical trial results to a target population. *Under review*.

[30] Lipid Research clinic Program. The Lipid Research Clinics Coronary Primary Prevention Trial results. I. Reduction in incidence of coronary heart disease. *JAMA 251*, 3 (1984), 351–364.

[31] Love, S. B., Brown, S., Weir, C. J., Harbron, C., Yap, C., Gaschler-Markefski, B., Matcham, J., Caffrey, L., McKevitt, C., Clive, S., et al. Embracing model-based designs for dose-finding trials. *British journal of cancer 117*, 3 (2017), 332.

[32] Mokdad, A. A., Xie, X.-J., Zhu, H., Gerber, D. E., and Heitjan, D. F. Statistical justification of expansion cohorts in phase 1 cancer trials. *Cancer 124*, 16 (2018), 3339–3345.

[33] Murthy, V. H., Krumholz, H. M., and Gross, C. P. Participation in cancer clinical trials: Race-, sex-, and age-based disparities. *Journal of the American Medical Association 291*, 22 (2004), 2720–2726.

[34] Neuenschwander, B., Branson, M., and Gsponer, T. Critical aspects of the bayesian approach to phase i cancer trials. *Statistics in medicine 27*, 13 (2008), 2420–2439.

[35] Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., Stuart, E. A., et al. Sensitivity analysis for an unobserved moderator in rct-to-target-population generalization of treatment effects. *The Annals of Applied Statistics 11*, 1 (2017), 225–247.

[36] NYC Department of City Planning. Decennial census, 2000.

[37] Odgaard-Jensen, J., Vist, G., Timmer, A., Kunz, R., Akl, E., Schünemann, H., Briel, M., Nordmann, A., Pregno, S., and Oxman, A. Randomisation to protect against selection bias in healthcare trials. *Cochrane database of systematic reviews*, 4 (2011).

[38] O'Muircheartaigh, C., and Hedges, L. V. Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 63*, 2 (2014), 195–210.

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

[40] Post, P. N., de Beer, H., and Guyatt, G. H. How to generalize efficacy results of randomized trials: recommendations based on a systematic review of possible approaches. *Journal of Evaluation in Clinical Practice 19*, 4 (2013), 638–643.

[41] Rogatko, A., Babb, J. S., Tighiouart, M., Khuri, F. R., and Hudes, G. New paradigm in dose-finding trials: patient-specific dosing and beyond phase i. *Clinical cancer research 11*, 15 (2005), 5342–5346.

[42] Rosenbaum, P. R., and Rubin, D. B. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association 79*, 387 (1984), 516–524.

[43] Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 5 (1974), 688–701.

[44] Rubin, D. B. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics 6*, 1 (1978), 34–58.

[45] Senn, S. *Statistical issues in drug development*, 2 ed. John Wiley & Sons, 2007.

[46] Simon, R., Rubinstein, L., Arbuck, S. G., Christian, M. C., Freidlin, B., and Collins, J. Accelerated titration designs for phase i clinical trials in oncology. *Journal of the National Cancer Institute 89*, 15 (1997), 1138–1147.

[47] Sommer, A., and Zeger, S. L. On estimating efficacy from clinical trials. *Statistics in Medicine 10*, 1 (1991), 45–52.

[48] Stuart, E. A., Bradshaw, C. P., and Leaf, P. J. Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science 16*, 3 (2015), 475–485.

[49] Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174*, 2 (2011), 369–386.

[50] Stuart, E. A., and Green, K. M. Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology 44*, 2 (2008), 395–406.

[51] Thall, P., and Lee, S.-J. Practical model-based dose-finding in phase i clinical trials: methods based on toxicity. *International Journal of Gynecologic Cancer 13*, 3 (2003), 251–261.

[52] Unger, J. M., Cook, E., Tai, E., and Bleyer, A. The role of clinical trial participation in cancer research: barriers, evidence, and strategies. *American Society of Clinical Oncology Educational Book 36* (2016), 185–198.

[53] Weiss, N. S. Generalizability of cancer clinical trial results. *Cancer 109*, 2 (2007), 341–341.

[54] Wheeler, G. M., Sweeting, M. J., and Mander, A. P. Aplusb: a web application for investigating a+ b designs for phase i cancer clinical trials. *PloS one 11*, 7 (2016), e0159026.