UNIVERSITY OF CALIFORNIA

Los Angeles

Evaluation of Clinical Trial Design Quality
Using Desirability Functions

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Biostatistics

by

Priscilla Kimberly Yen

2019

ABSTRACT OF THE DISSERTATION

Evaluation of Clinical Trial Design Quality Using Desirability Functions

by

Priscilla Kimberly Yen

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2019

Professor Weng Kee Wong, Chair

The design phase of a randomized controlled clinical trial is critical to its success. With many non-adaptive designs and an explosive number of adaptive designs introduced to the research community, the number of designs from which a statistician can select has the potential to be overwhelming. At times, a statistician may be uncertain how a newer adaptive design will perform in a particular setting of interest. While regulatory agencies have originally treated adaptive designs with resistance, recent years have seen more acceptance if there is extensive simulation work that shows good control of Type I error.

There are many adaptive designs, and it is important to understand and compare characteristics of competing designs before implementation. However, the overall lack of understanding of the performance of adaptive designs with regard to several design characteristics and the lack of an effective tool to measure overall design quality may have led to clinical trial statisticians implementing traditional designs rather than adopting more innovative methods. Yet adaptive designs have many appealing features that can benefit both the clinical trial sponsor, who funds the trial, and the clinical trial subjects. These strengths include early completion of a trial due to overwhelming efficacy and minimizing the number of subjects assigned to an inferior treatment arm.

The aim of this dissertation is to introduce methodology that provides statisticians and other clinical trial stakeholders with a tool that can measure the overall quality of a design and thereby facilitate comparison across competing designs. The methodology utilizes desirability functions to measure various statistical and non-statistical features that contribute to the quality of a design. Specifically, individual desirability functions evaluate a library of components including statistical considerations, such as treatment group size imbalance, probability of covariate imbalance, accidental bias, control for chronological bias, Type I error and power, and ethical considerations, such as minimizing the expected number of failures and total sample size needed in the whole trial. The proposed strategy is to compute an overall desirability score for each design, use it to rank the clinical trial designs of interest, and select the most relevant and efficient design for the trial's various objectives. To facilitate use of the proposed methodology, the project includes the development of an online interactive tool for the user to incorporate input before desirability functions are generated to help the user select the most appropriate design for the trial.

The dissertation of Priscilla Kimberly Yen is approved.

Thomas R Belin

Catherine Crespi-Chun

Hongquan Xu

Weng Kee Wong, Committee Chair

University of California, Los Angeles

2019

Contents

*

List of Figures

List of Tables

Notation

n_E	number of subjects in experimental arm (group E)
n_C	number of subjects in control arm (group C)
$n_E(j)$	number of subjects in experimental arm (group E) at the time of enrollment of the j^{th} patient
$n_C(j)$	number of subjects in control arm (group C) at the time of enrollment of the j^{th} patient
n	total sample size $= n_E + n_C$
Diff_j	$n_E(j) - n_C(j)$
φ	target allocation: the target proportion of subjects in the experimental arm E $= \frac{n_E}{n_E+n_C}$
ϕ	probability that subject j will be assigned to the experimental arm E
Y_{E_j}	response of subject j in experimental arm (group E)
Y_{C_j}	response of subject j in control arm (group C)
f_E	number of failures in experimental arm (group E)
f_C	number of failures in control arm (group C)
α	alpha-level: probability of rejecting a null hypothesis when in fact the null hypothesis is true
T_j	experimental arm indicator variable for subject j: 1 if in experimental arm E, 0 if in control arm C
iter	the number of iterations performed in a simulation, with each iteration completing one trial
d_i	individual desirability score for a realized value of the i^{th} characteristic $(i = 1, \cdots, m)$
k	a kurtosis parameter used in Harrington's desirability function
L	lower acceptable limit of a response's or characteristic's value
U	upper limit of a response's or characteristic's value
T	target value of a response's or characteristic's value (NTB variables only)
r	scale parameter for individual desirability functions (STB and LTB only)
r_1, r_2	scale parameters for individual desirability functions (NTB only)
m	the number of characteristics included in an overall desirability function
w_i	weight for characteristic i in calculation of D
D	overall desirability score for a design
p	the number of covariates that contribute to a response, used in genetic algorithms (Section 3.5)
θ	$(\mu_E, \mu_C, \sigma_E^2, \sigma_C^2)$
R	n_E/n_C
ρ	correlation between two treatment arms' responses
$\boldsymbol{p}$	(p_E, p_C) = probabilities of success in experimental and control arms
$\boldsymbol{\beta}$	(β_0, β_1) = (baseline outcome for the experimental arm, treatment effect for the experimental arm) (Details in Section 4.1.5)

Acknowledgments

This research has in part been supported by the AIDS Training Grant T32/AI-07370.

I would like to acknowledge my advisor, Dr. Weng Kee Wong, and my committee members for their time, thoughts, and review of this work. I owe my progression from a "I'll be done in two years" master's student to a committed doctoral student to the support of my family and the inspiration from my professors. Leaving behind a full-time job in finance to start anew at UCLA Biostatistics, I confirmed I was beginning down (or up?) a path that indeed aligned with my true interests thanks to Dr. Ron Brookmeyer's Introduction to Biostatistics class. Since then, the classes of Dr. Thomas Belin and Dr. Rob Weiss, as well as several mentors from industry, led me to wonder if perhaps I would want to pursue my PhD. I am forever grateful to the UCLA Department of Biostatistics for their support in my doctoral endeavours.

This journey would not have been possible without the love and support of my husband Simpson Wong, my parents Eugene and Kathy Yen, and my close friends and family.

Biographical Sketch

Education

Master of Science, Department of Biostatistics Sept 2012 - Dec 2014
University of California, Los Angeles

Bachelor of the Arts, Economics-Statistics Sept 2004 - May 2008
Columbia University in the City of New York

Professional Employment

Biostatistics Manager - Center for Design & Analysis
Amgen, Thousand Oaks Nov 2018 - present

Biostatistics Intern - US Medical Affairs Biometrics, Oncology Area
Genentech, South San Francisco May 2016 - Aug 2016

R&D Grad Intern - Global Statistical Programming, Bone Therapeutic Area
Amgen, Thousand Oaks Jun 2014 - Sept 2014
Amgen, Seattle Jun 2013 - Aug 2013

Research Associate - Web Analytics & Business Development Division
Kazaana, Inc., Menlo Park, CA Oct 2011 - Dec 2012

Data Operations Associate
Moody's Analytics, San Francisco, CA Jul 2008 - Oct 2011

Academic Employment

Teaching Assistant, UCLA Department of Biostatistics
100A - "Introduction To Biostatistics" (t-tests, confidence intervals, sampling) Mar 2018 - Jun 2018
100B - "Introduction to Biostatistics" (confounding, regression, odds ratio, survival) Jan 2018 - Mar 2018
203A - "Data Management & Statistical Computing" (SAS, R) Sept 2017 - Dec 2017
200B - "Biostatistics" Jan 2017 - Mar 2017

Graduate Student Researcher Mar 2017 - Jun 2017
Doctor Evidence, Santa Monica, CA

Graduate Student Researcher Nov 2013 - Jan 2017
UCLA AIDS Institute, Los Angeles, CA

Statistician - Department of Psychology Aug 2016
University of Stellenbosch, Stellenbosh, South Africa

College Academic Mentor Sept 2013 - Jun 2014
UCLA College Academic Counseling Department, Los Angeles, CA

Publications and Presentations

Bantjes J, Tomlinson M, Weiss RE, Yen PK, Goldstone D, Stewart J, Qondela T, Rabie S, Rotheram-Borus M-J. "Non-fatal suicidal behaviour, depression and poverty among young men living in low-resource communities in South Africa", BMC Public Health, 2018 Oct; 18:1195.

Epeldegui M, Magpantay L, Guo Y, Halec G, Cumberland W, Yen PK, Macatangay B, Margolick J, Rositch A, Wolinsky S, Martinez-Maza O, Hussain S: "A prospective study of serum microbial translocation biomarkers and risk of AIDS-related non-Hodgkin lymphoma", AIDS, 2018 Feb; epub ahead of print.

Grenon SM, Owens CD, Nosova EV, Hughes-Fulford M, Alley HF, Chong K, Perez S, Yen PK, Boscardin J, Hellmann J, Spite M, Conte MS. "Short-Term, High-Dose Fish Oil Supplementation Increases the Production of Omega-3 Fatty Acid-Derived Mediators in Patients With Peripheral Artery Disease (the OMEGA-PAD I Trial)", Journal of the American Heart Association, 2015 Aug; 4(8):e002034.

Nosova EV, Yen P, Chong KC, Alley HF, Stock EO, Quinn A, Hellmann J, Conte MS, Owens CD, Spite M, Grenon SM. "Short-term physical inactivity impairs vascular function", Journal of Surgical Research, 2014 Aug; 190(2):672-82.

Grenon SM, Owens CD, Alley H, Chong K, Yen PK, Harris W, Hughes-Fulford M, Conte MS. "n-3 Polyunsaturated fatty acids supplementation in peripheral artery disease: the OMEGA-PAD trial", Vascular Medicine, 2013 Oct; 18(5):263-74.

Awards

42nd Annual Lester Breslow Student Speaking Competition Winner	Apr 2016
Fielding School of Public Health Student Writing Competition Departmental Finalist	Jan 2016

Chapter 1

Randomization Procedures: A Review

The objective of this dissertation is to demonstrate the usefulness of desirability functions as a tool in the evaluation of clinical trial design quality. The motivation behind this research is driven by two reasons: first, newly defined adaptive designs in the literature are abundant but have yet to be commonly implemented in clinical trial practice, mainly due to uncertainty regarding the overall performance and quality of these new designs; second, the multitude of designs available today make a selection overwhelming.

An overview of the clinical trial designs considered in this dissertation is detailed in this chapter. Section 1.1 reviews nonadaptive designs - designs of trials with a predetermined sample size and treatment group allocation which do not change regardless of the data observed during the trial. Section 1.2 provides an introduction to response-adaptive randomization (RAR) - designs that seek to incorporate information from observed data during the trial to adjust sample size or treatment group allocations. Some popular target allocations of RAR are discussed.

Expanding upon RAR designs discussed in Section 1.2, Chapter 2 introduces a new target allocation that seeks to minimize total expected responses in two-arm trials when responses between the two arms are correlated. This new target allocation will be called R.corr throughout this work, and adds to the expansive list of designs a statistician might consider in the planning stages of a trial.

To evaluate design choices, it is helpful to review desirability functions and how they have been utilized in the biomedical research community thus far, which is the topic of Chapter 3. Chapter 4 follows with a framework implementing desirability functions for design quality evaluation. A few examples of implementation of the framework are provided. An online tool is available to the reader for exploration and utilization of this framework. Chapter 5 provides a more detailed case study of evaluating different designs that could be used in a clinical trial studying vertical transmission of HIV in pregnant women. The dissertation closes with an Epilogue summarizing the work and discussing future research topics.

* * *

The designs in this chapter touch on the surface of available designs for two-arm trials and are clearly not comprehensive; topics such as enrichment designs, platform designs, wedge designs, Bayesian designs, and designs for trials for personalized medicine are not discussed. Designs selected for review are commonly mentioned in educational textbooks on clinical trial procedures and statistical inference in randomized clinical trials. Specifically, the non-RAR designs of Section 1.1 are discussed as a foundation to understanding trial designs in Rosenberger, et. al's *Randomization in Clinical Trials: Theory and Practice* [55]. The textbook also touches on a few of the RAR designs in Section 1.2. A larger selection of the RAR designs in this work are inspired by their discussion in Menon, et. al's *Modern Approaches to Clinical Trials Using SAS: Classical, Adaptive, and Bayesian Methods* [61]. The purpose of this review is to familiarize the reader with such designs and their design characteristics to be discussed in Chapter 4. The reader is then free to extend the concepts in Chapter 4 to other types of designs, including designs with more than two treatment arms.

Although there are myriad designs widely discussed in the literature, most clinical trials of large pharmaceutical companies lean towards traditional designs such as complete randomization or permuted block designs. While the benefits of adaptive designs are not dismissed, trialists may be hesitant to utilize them due to their limited acceptance by regulatory agencies such as the Food and Drug Administration (FDA). The hesitance here is often attributed to uncertainty about control of the Type I error in adaptive designs, as well as other weaknesses that may offset their strengths, including concerns about biased treatment effect estimates in the presence of time trends. Chapter 4 is a useful contribution because it provides a framework to objectively assess strengths and weaknesses of various designs so that the overall quality of designs in regards to a specific research hypothesis may be better understood.

Throughout this thesis, we focus on 2-arm trials, denoting Treatment E as the experimental arm, and Treatment C as the control arm, and we assume we have resources to recruit n patients. A set of treatment assignments for n patients is $T_1, ..., T_n$, where $T_j = 1$ when patient j is assigned to experimental arm E, and $T_j = 0$ when patient j is assigned to control arm C. The probability of being assigned to the experimental arm E is denoted by $Pr(T_j = 1) = E(T_j)$. In this chapter, we denote the difference in sample size between experimental arm E and control arm C when subject j is enrolled with $\text{Diff}_j = n_E(j) - n_C(j)$.

1.1 Nonadaptive Designs

Nonadaptive designs allocate subjects to a specific treatment arm with a probability that is independent of how other subjects have responded so far in the trial, as well as of the current subject's baseline characteristics. Because these design do not *adapt* to prior subject performance or current subject's covariates, they are called "nonadaptive" designs. Amongst nonadaptive designs, we discuss the traditional Complete Randomization Design, Forced Balance Designs, and Biased Coin Designs.

Complete Randomization Design (CRD)

In Complete Randomization Design (CRD), each patient is enrolled into either treatment arm E or control arm C with probability 1/2. There are no restrictions imposed upon this design.

An advantage of CRD is the reduction of certain types of biases: such as selection bias - since it is equally likely to guess the next treatment assignment correctly or incorrectly, and biases due to covariate imbalance - since the completely random assignment of treatments is expected to lead to a balance in covariates between treatment arms.

One disadvantage of CRD is imbalance between treatment group size. While different treatment group sizes does not bias the estimate of the treatment effect, larger imbalances lead to less precision of the estimate, and hence, a decrease in power. A second disadvantage is that it is unable to address ethical concerns: as more information is learned about the two treatment arms, complete randomization will ignore this information and still assign patients to either treatment arm with equal probabilities. Consequently, the design does not incorporate concern for the patient's well-being and does not increase the probability of being assigned to what is considered the superior treatment arm at the time.

1.1.1 Forced Balance Designs

Truncated Binomial Design (TBD)

The Truncated Binomial Design (TBD) (Blackwell and Hodges, 1957) is a forced balance procedure, meaning that exactly half of n patients will be assigned to each treatment arm. In this allocation rule, complete randomization is performed until one treatment arm contains half of the pre-determined sample size; subsequently, all remaining patients will receive the other treatment.

Let $\mathcal{F}_n = T_1, ..., T_n$ be a set of treatment assignments for n stages of the randomization process. Then, the truncated binomial design allocation rule is defined by:

$$E(T_j|\mathcal{F}_{j-1}) = \frac{1}{2}, \text{if } max(n_E(j-1), n_C(j-1)) < \frac{n}{2}$$

$$= 0, \text{if } n_E(j-1) = \frac{n}{2}$$

$$= 1, \text{if } n_C(j-1) = \frac{n}{2}.$$

Note that $Pr(T_n = 1) = E(T_n)$. While the truncated binomial design offers balance treatment arm sizes, when one treatment arm is considered "full", it is clear that the remaining patients will be forced into the remaining treatment arm, resulting in selection bias and high risk of covariate imbalances.

Random Allocation Rule (RAR)

The Random Allocation Rule (RAR) is also a forced balance procedure, with

$$E(T_j|\mathcal{F}_{j-1}) = \frac{\frac{n}{2} - n_E(j-1)}{n-(j-1)}, j = 2, ..., n,$$

and $E(T_1) = 1/2$.

One can think of this allocation rule in terms of an urn model. One samples from an urn with n/2 balls for experimental arm E, and n/2 balls for control group C, without replacement. While the RAR guarantees treatment group size balance, it shares the TBD's weakness of having 100% predictability of treatment assignments once n/2 patients have been assigned to one of the treatment groups. A second weakness is its susceptibility to covariate imbalance.

Permuted Block Design (PBD)

While Truncated Binomial Design guarantees balance in treatment group size at the end of the trial, it does not guarantee balance at several time points during the trial. In fact, all two of three designs previously discussed are prone to severe treatment size imbalance at some point during the trial.

In order to avoid severe treatment size imbalance during the entire course of a trial, clinical trialists often use "blocks". Forced balance randomization within blocks is used in order to ensure balance at the end of each block. Specifically, in the Permuted Block Design (PBD) (Zelen, 1974), there are M blocks of size B, where $B = n/M$. Each block is filled using a forced balanced procedure (e.g. Random Allocation Rule, Truncated Binomial Design), so that there are M occurrences of balanced allocation during the course of the trial. The maximum imbalance at any time point is then half a block size, $B/2$.

Let R_j define the position patient i takes within his block. If we fill blocks using RAR, the allocation rule is:

$$E(T_j|\mathcal{F}_{j-1}, B, R_j) = \frac{\frac{B}{2} - \sum_{l=j+1-R_j}^{j-1} T_l}{B - R_j + 1}.$$

Ensuring balance consistently throughout a trial is important when patient's outcomes or covariates follow a time trend throughout the trial. If there are time-heterogeneous covariates, allowing large treatment size imbalances during a trial while using CRD, RAR, or TBD could result in significant covariate imbalances. Similarly, if the trial's outcome of interest follows a time trend, large treatment size imbalances would result in uncertainty regarding whether an observed treatment effect was in fact due to treatment itself or due to the treatment size imbalance. These are examples of chronological bias, which will be detailed more in Section 4.1.5.

Random Block Design (RBD.RAR, RBD.TBD)

Just as forced balance procedures are subject to selection bias, so is the permuted block design, which implements forced balance procedures within each "block" of patients. This is because, if the block size is known, those in charge of enrolling patients may realize at one point that the probability of being assigned to a specific treatment is guaranteed.

The Random Block Design protects against this risk, since block sizes are randomly selected from a discrete uniform distribution. Let B_{max} be the maximum imbalance of number of subjects in the two arms, which is half of the largest block. The different block sizes, picked at random with probability $1/B_{max}$ after the fulfillment of a single block, are then 2, 4, 6, ..., $2B_{max}$. Let B_j be the block size of the block with the j^{th} patient. Let R_j define the position patient j takes within his block, ranging from 1,...,B_j. Each block can be filled with any forced balance procedure. If we fill each block using RAR, the allocation rule is:

$$E(T_j|\mathcal{F}_{j-1}, B_j, R_j) = \frac{\frac{B_j}{2} - \sum_{l=j+1-R_j}^{j-1} T_l}{B_j - R_j + 1}.$$

In this work, we refer to random block design filling blocks with random allocation rule as RBD.RAR, and random block design filling blocks with truncated binomial design as RBD.TBD. The number of possible block sizes fluctuates depending on B_{max}, with larger values of B_{max} allowing for more potential block sizes, yet also exposing the trial to the risk of a treatment size imbalance as large as B_{max}. One way to address this is to use only permutations of block sizes that sum to exactly a pre-determined total sample size n.

1.1.2 Biased Coin Designs

Biased coin designs aim to obtain approximately equal allocation while still allocating subjects to treatments with some randomness. They are different from response-adaptive designs discussed in the next Section (Section 1.2), because biased coin designs consider only the treatment allocation history, and response-adaptive designs also take into account patient responses or baseline covariates. Atkinson (2014) provides reviews of various biased-coin designs and their ability to achieve treatment group size balance [6]. This subsection introduces a number of biased coin designs; Chapter 4 evaluates the first five as candidate designs.

Efron's Biased Coin Design (BCD_p)

The Biased Coin Design (Efron, 1971) seeks to provide approximate balance of treatment assignments whenever the trial is stopped while still providing randomization to reduce biases. This is achieved by allocating patients to the underrepresented treatment group with a higher, fixed probability. The allocation rule is defined as

$$E(T_j|\mathcal{F}_{j-1}) = \begin{cases} \frac{1}{2}, & \text{if } |\text{Diff}_{j-1}| = 0, \\ p, & \text{if } \text{Diff}_{j-1} < 0, \\ 1-p, & \text{if } \text{Diff}_{j-1} > 0, \end{cases} \tag{1.1}$$

where $0.5 < p <= 1$, and Diff_j is defined just before Subsection 1.1. Clearly, when p = 1/2, we have complete randomization with the restriction of a maximal imbalance of n/2, and when p = 1, Efron's biased coin design simplifies to a permuted block design with a block size of 2, so that every other patient's allocation assignment is deterministic, with a maximal imbalance of 0 if n is even. The parameter p thus represents a trade-off between balance and predictability. Efron's original paper states: "The value p = 2/3, which is the author's personal favourite, will be seen to yield generally good designs..."

Big Stick Design (BSD)

The Big Stick Design (BSD) allows a degree of imbalance up to a magnitude given by a fixed imbalance tolerance parameter b. The allocation rule is given by:

$$E(T_j|\mathcal{F}_{j-1}) = \begin{cases} \frac{1}{2}, & \text{if } |\text{Diff}_{j-1}| < b, \\ 0, & \text{if } \text{Diff}_{j-1} = b, \\ 1, & \text{if } \text{Diff}_{j-1} = -b \end{cases} \tag{1.2}$$

Big Stick Design (proportion) (proportionBSD)

The Big Stick Design with Maximum Proportionate Degree of Imbalance replaces the absolute difference used in the Big Stick Design with an acceptable degree of imbalance, $\text{Diff}_{j-1}/(j-1)$. The allocation rule, then, is:

$$E(T_j|\mathcal{F}_{j-1}) = \begin{cases} \frac{1}{2}, & \text{if } \text{Diff}_{j-1}/(j-1) < \text{prop}, \\ 0, & \text{if } \text{Diff}_{j-1}/(j-1) = \text{prop}, \\ 1, & \text{if } \text{Diff}_{j-1}/(j-1) = -\text{prop}, \end{cases} \tag{1.3}$$

where prop is a pre-defined acceptable degree of imbalance.

Biased Coin Design with Imbalance Intolerance (BCDII(p))

The Biased Coin Design with Imbalance Intolerance (BCDII(p)) combines the concepts of the big stick design and Efron's biased coin design [19]. The allocation rule is defined by:

$$E(T_j|\mathcal{F}_{j-1}) = \begin{cases} \frac{1}{2}, & \text{if } \text{Diff}_{j-1} = 0, \\ 0, & \text{if } \text{Diff}_{j-1} = b, \\ 1, & \text{if } \text{Diff}_{j-1} = -b, \\ p, & \text{if } 0 < \text{Diff}_{j-1} < b, \\ 1-p, & \text{if } -c < \text{Diff}_{j-1} < 0 \end{cases} \tag{1.4}$$

This allocation rule results in a random walk on the space of 0,...,b, with reflecting barriers 0 and b.

Accelerated Biased Coin Design (ABCD(a))

The Accelerated Biased Coin Design (ABCD(a)) is a bigger umbrella of biased coin designs with a parameter a. It contains the big stick design, Efron's biased coin design, and biased coin design with imbalance intolerance as special cases. Let F be a function that maps integers to [0,1] such that $F(x)$ is decreasing, and $F(-x) = 1 - F(x)$. The allocation rule is then:

$$E(T_j|\mathcal{F}_{j-1}) = F(\text{Diff}_{j-1}),$$

where

$$F_a(x) = \begin{cases} \frac{|x|^a}{|x|^a+1}, & \text{if } x \leq -1, \\ \frac{1}{2}, & \text{if } x = 0, \\ \frac{1}{|x|^a+1}, & \text{if } x \geq 1. \end{cases} \tag{1.5}$$

The parameter a controls the degree of randomness, with $a = 0$ equating complete randomization [4]. As $a \to \infty$, the ABCD is equivalent to the Big Stick Design with $b = 2$. The design is called 'accelerated' because the parameter a exponentially weights the imbalance Diff_{j-1}.

Wei's Urn Design

The urn design is inspired by an urn that contains α balls representing subjects to be assigned to treatment group E, and α balls representing subjects to be assigned to control arm A. When a subject is randomized, a ball is drawn and subsequently replaced. If the ball drawn assigns a subject to group E, then β balls for group C are added to the urn, thus increasing the probability that the next subject will be assigned to the control arm C. Similarly, if the ball drawn assigns a subject to group C, then β balls for group E are added to the urn, thus increasing the probability that the next subject will be assigned to the experimental arm E. Thus, at any point, the urn composition is skewed so that the probability of a subject being assigned to the underrepresented arm is higher. The allocation rule is

$$E(T_1|F_0) = 1/2.$$

$$E(T_j|F_{j-1}) = \frac{\alpha + \beta n_C(j-1)}{2\alpha + \beta(j-1)}, j \geq 2.$$

Note that this is different than Efron's biased coin design and Big Stick Design because the degree of the probabilities of assignment under Wei's design alter according to the degree of imbalance [89, 90].

Generalized Biased Coin Design (GBCD)

Smith (1984) introduced a more general biased coin design, with an allocation rule given by

$$E(T_j|F_{j-1}) = \frac{n_C^\gamma}{n_E^\gamma + n_C^\gamma},$$

with γ controlling the randomness of the design. When $\gamma = 0$, the GBCD reduces to complete randomization. On the other hand $\gamma = 1$ reduces to Wei's urn design. Smith recommends $\gamma = 5$ in his paper for its ability to perform well with regards to degree of balance achieved, selection bias, accidental bias, and statistical inference of post-trial results.

1.2 Response-Adaptive Randomization

Response-Adaptive Randomization (RAR) designs adapt to the cumulative responses observed at a pre-specified period in the study and may increase or decrease the probability a subject is assigned to a given treatment arm. RAR designs target an allocation proportion, which depends on pre-stated objectives. The next subsection provides an overview of target allocation schemes frequently discussed in literature.

1.2.1 Target Allocations

Target allocation is defined as the ideal proportion of subjects placed in the experimental arm E: $n_E/(n_E + n_C)$. The ideal proportion varies depending on one's objectives. For example, urn models target an allocation proportion inversely proportional to the corresponding failure rate in binary response trials. Neyman allocation seeks to maximize power for a given sample size. RSIHR allocation, named after the initials of the authors on the original paper (Rosenberger, Stallard, Ivanova, Harper, and Ricks), seeks to minimize the expected number of failures of a trial with a given power [72]. Biswas and Mandal (BM) allocation is a generalization of optimal allocation to normal responses; henceforth we term their proposed allocation as BM allocation [13]. While the aforementioned allocations are optimal relative to some objective, Bandyopadhyay and Biswas (BB) allocation assigns treatment according to a mapping of the current difference in means [7].

These target allocations schemes are summarized in Table B.2.

Objective	Allocation Name	Binary	Continuous Normal
1. Maximize power for a fixed sample size	Neyman	$\frac{\sqrt{p_E q_E}}{\sqrt{p_E q_E}+\sqrt{p_C q_C}}$	$\frac{\sigma_E}{\sigma_E+\sigma_C}$
2. Minimize expected number of treatment failures for a fixed power	RSIHR	$\frac{\sqrt{p_E}}{\sqrt{p_E}+\sqrt{p_C}}$	$\frac{\sqrt{\mu_C}\sigma_E}{\sqrt{\mu_C}\sigma_E+\sqrt{\mu_E}\sigma_C}$
3. Minimize treatment failures and ensure fewer patients are allocated to inferior treatment	RSIHR2	NA	$\frac{\sqrt{\mu_C}\sigma_E}{\sqrt{\mu_C}\sigma_E+\sqrt{\mu_E}\sigma_C}$ if ($\mu_E < \mu_C$ and $\sigma_E\sqrt{\mu_C}/\sigma_C\sqrt{\mu_E}) > 1$ or ($\mu_E > \mu_C$ and $\sigma_E\sqrt{\mu_C}/\sigma_C\sqrt{\mu_E}) < 1$; 1/2 otherwise
4. Urn model (good for lowering expected number of treatment failures when $p_E + p_C > 1$)	Urn	$\frac{q_C}{q_E+q_C}$	NA
5. Minimize patients with response greater than c	Biswas Mandal (BM)	NA	$\frac{\sqrt{\Phi(\frac{\mu_C-c}{\sigma_C})}\sigma_E}{\sqrt{\Phi(\frac{\mu_C-c}{\sigma_C})}\sigma_E+\sqrt{\Phi(\frac{\mu_E-c}{\sigma_E})}\sigma_C}$
6. Not formally defined	Bandyopadhyay Biswas (BB)	NA	$\Phi(\frac{\mu_C-\mu_E}{T})$
7. Minimize the maximum eigenvalue of the inverse of Fisher's information (E-optimality)	Baldi Antognini Giovagnoli (Baldi)	see text	NA

Table 1.1: Summary of Allocations Targeted By RAR Designs.

Neyman Allocation

The Neyman allocation maximizes power for a given sample size n and fixed probabilities of success in the treatment arms [72]. The Neyman allocation can be applied to either binary or continuous responses [96]. The derivation in the binary case is presented here:

Consider a two-arm clinical trial with a binary response. Let p_E be the probability of success in experimental arm E, and p_C the probability of success in control group C. Then, q_E and q_C are the probabilities of failure for treatment arms A and B, respectively. Let there be a fixed n_E patients in group E and n_C patients in C, with $n_E + n_C = n$. To test whether the probability of success in the two treatment groups is equal, we test the hypothesis:

$$H_0 : p_E - p_C = 0 \text{ versus } H_1 : p_E - p_C \neq 0.$$

We apply the Wald test for a user-specified Type I error rate using the test statistic

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\frac{\hat{p}_A \hat{q}_A}{n_E} + \frac{\hat{p}_B \hat{q}_B}{n_C}}}.$$

Rather than fixing the sample proportions as $n_E = n_C = n/2$ and finding the minimum sample size n that achieves the desired power, we can instead fix the variance of the test under the alternative hypothesis, and find the allocation proportion that minimizes the total sample size. The equivalent question is: *for fixed sample size, what allocation maximizes power?*

Mathematically, the general optimization problem has the form:

$$\begin{aligned} &\min_{n_E, n_C} w_E n_E + w_C n_C \\ &\text{subject to } Var(p_E - p_C) = K, \end{aligned} \tag{1.6}$$

where w_E and w_C are positive weights, and $\boldsymbol{p}^T = (p_E, p_C)$. For the problem at hand, we seek to fix the variance and minimize the total sample size, so our weights w_E and w_C are both equal to 1.

Let φ be the proportion of patients to be assigned to experimental arm E. Then $n_E = \varphi n$, $n_C = (1-\varphi)n$, and the variance of the estimated difference in probabilities of success $p_E - p_C$ is

$$Var(\hat{p_E} - \hat{p_C}) = \frac{p_E q_E}{\varphi n} + \frac{p_C q_C}{(1-\varphi)n} := K,$$

and hence

$$n = \frac{p_E q_E}{\varphi K} + \frac{p_C q_C}{(1-\varphi)K}. \tag{1.7}$$

We minimize Eq. 1.7 with respect to φ by taking the derivative, setting equal to zero, and solving for φ. This gives

$$\frac{dn}{d\varphi} = -\frac{p_Eq_E}{\varphi^2} + \frac{p_Cq_C{}^2}{1-\varphi} = 0$$

and

$$\varphi = \frac{\sqrt{p_Eq_E}}{\sqrt{p_Eq_E} + \sqrt{p_Cq_C}}.$$

This is known as the Neyman allocation. A weakness of this allocation is that when the success probabilities on both treatment groups are high ($p_E + p_C > 1$), more subjects are assigned to the weaker treatment arm.

RSIHR Allocation

While one objective could be to minimize the total sample size in a trial while still achieving sufficient power, a second objective could be instead to minimize the total number of treatment failures. This objective equates to $w_A = q_E$ and $w_B = q_C$ in Eq. 1.6. RSIHR allocation (named after the initials of the authors on the original paper) seeks to minimize treatment failures for a fixed power [72], and the allocation is given by:

$$\varphi = \frac{\sqrt{p_E}}{\sqrt{p_E} + \sqrt{p_C}}.$$

In the continuous case, φ is derived from solving the optimization problem:

$$\begin{aligned} &\min_{\frac{n_E}{n_C}} \mu_E n_E + \mu_C n_C \\ &s.t. \frac{\sigma_E^2}{n_E} + \frac{\sigma_C^2}{n_C} = K. \end{aligned}$$

A direct calculation yields the optimal proportion of subjects in the experimental arm to be

$$\varphi = \frac{\sqrt{\mu_C}\sigma_E}{\sqrt{\mu_C}\sigma_E + \sqrt{\mu_E}\sigma_C}.$$

However, when $\mu_E < \mu_C$, it is possible for $\frac{n_E}{n_C}$ to be less than 1/2. This shows that while power is maximized for a fixed expected number of treatment failures, this target allocation has the potential to allocate more patients to the inferior treatment. RSIHR2 seeks to remove this ethical flaw by modifying the allocation [96].

Urn Allocation

In trials with binary responses, urn allocation can be used when the probability of success is high in both the experimental and the control arms, specifically, when $p_E + p_C > 1$. The target allocation is

$$\varphi = \frac{q_E}{q_E + q_C}.$$

Urn models such as the randomized play-the-winner (RPW) (Wei et al., 1978) and drop-the-loser (DL) rule (Ivanova, 2003) are procedures that target the Urn allocation, which is only a property of the procedure rather than the solution to a specific optimality problem.

Biswas and Mandal Allocation (BM)

Biswas and Mandal (2004) sought to extend the concept of minimizing failures in trials with binary outcomes to trials with normal responses. The resulting allocation assumes that smaller responses are better and minimizes the total number of patients with response greater than c, thereby minimizing the number of failures as defined by a threshold [13]. The target allocation is

$$\varphi = \frac{\sqrt{\Phi(\frac{\mu_C - c}{\sigma_C})}\sigma_E}{\sqrt{\Phi(\frac{\mu_C - c}{\sigma_C})}\sigma_E + \sqrt{\Phi(\frac{\mu_E - c}{\sigma_E})}\sigma_C}.$$

Bandyopadhyay and Biswas Allocation (BB)

Bandyopadhyay and Biswas (2001) proposed a target allocation that does not seek to optimize any formal objective property. S is a scaling factor or “tuning constant”, with larger values of S resulting in higher proportions of allocation to the better treatment, but also higher variability in the allocation proportion. The values 1, 2, and 3 are considered in the original paper. The authors recommend to set $S = 2$ [7]. The target allocation is

$$\varphi = \Phi(\frac{\mu_C - \mu_E}{S}).$$

Baldi Antognini and Giovagnoli (Baldi)

Baldi Antognini and Giovagnoli (2010) proposed a target allocation that has both ethical and inferential aims [5]. They seek to minimize the maximum eigenvalue of the inverse of Fisher’s information (E-optimality). Consider the compound criterion which combines ethical and inferential objectives:

$$\Phi_w(\varphi) = w\left(\frac{\psi_E(\varphi)}{\psi_E^*}\right) + (1-w)\left(\frac{\psi_I(\varphi)}{\psi_I^*}\right),$$

where $w \in (0,1)$ is a user-defined weight for importance of ethics, $1 - w$ is the weight given to inference, and $\psi_E(\varphi) = q_E\varphi + q_C(1-\varphi)$ is the expected proportion of treatment failures, $\psi_E^* = min(q_E, q_C)$, $\psi_I(\varphi) = p_E q_E/\varphi + p_C q_C/(1-\varphi)$ is the variance of the estimated treatment difference, and $\psi_I^* = (\sqrt{p_E q_E} + \sqrt{p_C q_C})^2$ is the minimum value of $\psi_I(\varphi)$ for $\varphi \in (0,1)$. The goal is to minimize the compound criterion. We can see, then, w places more importance on minimizing $\psi_E(\varphi)$, the expected proportion of failures, and $(1-w)$ places more importance on $\psi_I(\varphi)$, which minimizes the variance of the estimated treatment difference.

The target allocation φ is the solution in (0,1) of the following equation

$$\frac{w}{1-w}\frac{p_E - p_C}{\min(q_E, q_C)}\Big(\frac{\sqrt{p_C q_C}}{\sqrt{p_E q_E}}+1\Big)^2 = \frac{(\frac{\sqrt{p_C q_C}}{\sqrt{p_E q_E}}-1)\varphi^2 + 2\varphi - 1}{(\varphi(1-\varphi))^2}.$$

The examples in this dissertation will use $w = 1/2$, giving equal weight to ethics and inference.

1.2.2 Response-Adaptive Randomization (RAR) Designs

Several randomization procedures have been proposed that can target a specific allocation such as those discussed in the previous subsection. These randomization procedures utilize an allocation function which calculates the probability that the next subject is to be assigned to the treatment arm. One desirable characteristic of an RAR procedure is its ability to attain the target allocation by the end of subject enrollment in a trial. Some examples of response-adaptive randomization (RAR) designs are discussed in this section.

Melfi & Page, 2000 (MP)

Let $Y_1, \ldots, Y_n$ be subject responses. If $T_j = 1$, Y_j is normally distributed with mean μ_E and variance σ_E^2. If $T_j = 0$, Y_j is normally distributed with mean μ_C and variance σ_C^2. A simple allocation function is to simply set the probability of being assigned to the experimental arm to be the value of the target allocation $\varphi(\hat{\theta}_j)_{\text{target}}$, where the hat indicates that the target allocation is estimated after j responses.

$$\phi = \varphi(\hat{\theta}_j)_{\text{target}}$$

Melfi et al. (2000) showed that $n_E/n \to \varphi(\theta)$ almost surely [60]. However, this function that directly sets the probability of being assigned to the experimental arm to be directly equal to the target allocation proportion induces high variability. Extra variance in the treatment group sizes n_E and n_C can also negatively impact the optimal properties of $\varphi(\theta)$ [55].

Eisele & Woodroofe, 1995 (EW1995)

Eisele & Woodroofe (1995) presented a response-adaptive randomization (RAR) procedure [29]. Assume the same response model as presented in MP design above. Let

$$g(x, y) = [1 - (\frac{1}{y} - 1)x],$$

where $x \in [0, 1]$ and $y \in [0, 1]$. Then Eisele & Woodroofe's procedure is defined by:

$$\phi = g\bigg(\frac{n_E(j-1)}{j-1}, \varphi(\hat{\theta_j})_{\text{target}}\bigg).$$

Doubly-Biased Coin Design (DBCD)

The Doubly Biased Coin Design (DBCD) is a RAR procedure that obtained its name from consideration of both the proportion of enrolled patients assigned to each treatment arm and the estimate of the target allocation proportion [47]. Biased coin designs are able to reduce experimenter/selection bias.

The procedure aims to fulfill the goal of allocating n_E patients to treatment E, such that $\frac{n_E}{n_E+n_C}$ equals the target allocation proportion. Treatment E is assigned, then, with a probability less than the current maximum likelihood estimate (MLE) of the target proportion when the observed proportion is larger than this estimate. Similarly, Treatment E is assigned with a probability greater than the current MLE of the target proportion when the observed proportion is larger than this estimate.

One strength of the DBCD is its reduction in variability of treatment assignments, which it does by reducing the distance between $n_E(j)$ and $\varphi(\hat{\theta_j})$. Specifically, let allocation function $g(x, y)$ be defined on $[0, 1] \times [0, 1]$, where

$$g(x,y) = \frac{y(y/x)^\gamma}{y(y/x)^\gamma + (1-y)((1-y)/(1-x))^\gamma}, \tag{1.8}$$

and the DBCD assignment rule is defined by:

$$\phi = g\bigg(\frac{n_E(j-1)}{j-1}, \varphi(\hat{\theta_j})_{\text{target}}\bigg). \tag{1.9}$$

Note that if the procedure seeks to consistently target $\varphi(\theta) = 0.5$, the design is no longer response-adaptive, but reduces to Smith's generalized biased coin design (GBCD) as discussed in Section 1.1.2. Rather than setting $\varphi(\theta)$ to be constant, the target allocation is recalculated for each enrolling patient using estimates of the means and variances of the previously enrolled patients. The nonnegative parameter γ determines the degree of randomness, with $\gamma = 0$ being the most random due to ϕ simplifying to $\varphi(\hat{\theta_j})_{\text{target}}$. When $\gamma \to \infty$, DBCD becomes an almost deterministic procedure [47]. In this dissertation, we use $\gamma = 2$, which was recommended by Hu et al. (2004) as a suitable trade-off between reducing variability and maintaining sufficient randomness.

Below is a step-by-step algorithm for constructing a DBCD:

- Step 1. For given error rates Type I (α) and Type II (β), and nominal values of $\mu_E, \mu_C, \sigma_E, \sigma_C$, determine the total sample size n, where $n = n_E + n_C$, and

$$n_E = n_C = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2(\sigma_E^2 + \sigma_C^2)}{(\mu_E - \mu_C)^2}.$$

- Step 2. Let m_0 be a user-selected block size for an initial $2m_0$ subjects. These first $2m_0$ subjects are enrolled using permuted block design during a "run-in period", where we do not yet want DBCD to be implemented due to the little amount of data accrued. Specifically, enroll $2m_0$ subjects, with the probability of each subject j being placed in treatment E equaling $\frac{m_0 - n_E(j-1)}{2m_0-(j-1)}$, where $j = 1 \dots 2m_0$, and $n_E(j-1)$ is the number of patients enrolled in treatment arm E at the time of the $(j-1)^{st}$ patient. The purpose of this step is to gather enough patients into each treatment group so that estimation of $\hat{\mu}_E, \hat{\mu}_C, \hat{\sigma_E}^2, \hat{\sigma_C}^2$ can be done in the next step.

- Step 3. Estimate $\hat{\mu}_E, \hat{\mu}_C, \hat{\sigma_E}^2, \hat{\sigma_C}^2$ from the available data.

- Step 4. Let $\boldsymbol{T_j} = \mathbb{1}\boldsymbol{(}$subject j in experimental arm E). Then, the probability that subject j is enrolled into experimental arm E can also be denoted as the expected value of T_j, which is calculated using the assignment rule in Equation 1.9.

- Step 5. Generate a value U from the Uniform[0,1] distribution. If $U < [E(T_j|F_(j-1)]$, assign subject j to treatment arm E. Otherwise, assign subject j to treatment arm C.

- Step 6. Repeat Steps 3 through 5 until n total subjects have been enrolled.

Sequential Maximum Likelihood Estimation Design (SMLE)

The Sequential Maximum Likelihood Estimation (SMLE) Design sets treatment randomization probabilities to be equal to the current estimates of the target allocation proportions. It is equivalent to DBCD with $\gamma = 0$. This design can lead to a modest reduction in treatment failures with minimal loss in power relative to equal randomization designs [55]. However, SMLE has also been shown to be quite variable, with potential negative effects on power [61].

Efficient Randomized-Adaptive Design (ERADE)

Efficient Randomized-Adaptive Design (ERADE) is an extension of the biased-coin design which targets equal allocation [48]; ERADE is equivalent to DBCD with $\gamma \to \infty$. ERADE is a RAR procedure that can target any pre-specified allocation proportion, while still preserving allocation randomness and boasting minimal variability. The theoretical properties of ERADE echo those of DBCD: both resulting sample proportions (n_E/n) and estimators are strongly consistent and asymptotically normal.

$$\phi = \begin{cases} \delta\hat{\varphi}_(j-1) & \text{if } n_E(j-1)/n > \hat{\varphi}_(j-1) \\ \hat{\varphi}_(j-1) & \text{if } n_E(j-1)/n = \hat{\varphi}_(j-1) \\ 1-\delta(1-\hat{\varphi}_(j-1)) & \text{if } n_E(j-1)/n < \hat{\varphi}_(j-1) \end{cases}, \tag{1.10}$$

where $0 \leq \delta \leq 1$ is a user-specified parameter that controls the degree of randomness of the randomization procedure. When $\delta = 0$, the procedure is the most deterministic; when $\delta = 1$, the procedure is most random. Hu et al. recommend $0.4 \leq \delta \leq 0.7$ [48]. In this dissertation, we use $\delta = 0.4$.

The discrete property of ERADE differentiates it from DBCD, resulting in less variability. In fact, ERADE is an asymptotically best procedure, attaining minimum variance. In the binary case, the minimum variance is

$$\frac{1}{4(\sqrt{p_E(1-p_E)}+\sqrt{p_C(1-p_C)})^3}\left(\frac{p_C(1-p_C)((1-p_E)-p_E)^2}{\sqrt{p_E(1-p_E)}}+\frac{p_E(1-p_E)((1-p_C)-p_C)^2}{\sqrt{p_C(1-p_C)}}\right)$$

as shown by Hu et al. (2009) [48].

1.2.3 Inference in Response-Adaptive Designs

The likelihoods for non RAR versus RAR designs are different. In restricted randomization (such as in biased coin designs), n_E can be random, yet it is still ancillary due to independence from observed responses. On the other hand, in RAR, n_E depends on observed responses and is thus no longer ancillary [55]. A "guiding principle" is that for RAR to be practical, standard inferential tests should be able to be utilized at the end of a trial [46]. The impact of n_E's dependence on observed responses is seen in the bias of the treatment effect estimate in response-adaptive designs. For example, Coad et al. showed that in a trial with binary responses, the bias in estimating the probability of success p_E is given by

$$E(\hat{p}_E - p_E) = p_E(1-p_E)\frac{\partial}{\partial p_E}E\left(\frac{1}{n_E(n)}\right).$$

Let $Y_1, \ldots, Y_n$ be independently and identically distributed, with $Y_{1k} \sim f_k(l, \theta_k), k \in (E, C)$, where $\theta_k \in$ parameter space Θ_k. For binary responses, $f_E(\cdot, \theta_E)$ is Bernoulli(p_E), $f_C(\cdot, \theta_C)$ is Bernoulli(p_C). For normal responses, $f_E(\cdot, \theta_E)$ is (μ_E, σ_E) and $f_C(\cdot, \theta_C)$ is (μ_C, σ_C). Assume the following three regularity conditions hold true:

- the parameter space Θ_k is an open subset of R^2,

- the distributions $f_E(\cdot, \boldsymbol{\theta}_E), f_C(\cdot, \boldsymbol{\theta}_C)$ follow an exponential family

- the limiting allocation $\boldsymbol{\varphi}(\boldsymbol{\theta}) = (\varphi_E(\boldsymbol{\theta}), \varphi_C(\boldsymbol{\theta}))$ has $\frac{n_k}{n} \to \varphi_k(\boldsymbol{\theta})$ for $k \in E, C$ almost surely [46].

Then $\hat{\boldsymbol{\theta}}$ is strongly consistent for $\boldsymbol{\theta}$ and $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$, where $\boldsymbol{I}$ is the Fisher's information matrix $\boldsymbol{I}(\boldsymbol{\theta}) = \operatorname{diag}(\varphi_E(\boldsymbol{\theta})\boldsymbol{I}_E(\boldsymbol{\theta}_E), \varphi_C(\boldsymbol{\theta})\boldsymbol{I}_C(\boldsymbol{\theta}_C))$ and $\boldsymbol{I}_k(\boldsymbol{\theta}_k) = -E\Bigg(\frac{\partial^2 \log f_k(Y_{1k}, \boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k^2}\Bigg)$ is the Fisher's information for a single observation on treatment $k \in E, C$.

For example, for the DBCD that targets Neyman allocation, we have

$$\sqrt{n}\Bigg(\begin{bmatrix}\hat{\mu}_E \\ \hat{\mu}_C\end{bmatrix} - \begin{bmatrix}\mu_E \\ \mu_C\end{bmatrix}\Bigg) \to N\Bigg(\mathbf{0}, \begin{bmatrix}\sigma_E(\sigma_E + \sigma_C) & 0 \\ 0 & \sigma_C(\sigma_E + \sigma_C)\end{bmatrix}\Bigg)$$

in distribution. The bias of different designs is one of many components of a design that are evaluated during overall trial quality assessment in Chapter 4.

The variability of a randomization procedure is directly tied with its power [46]. To show this, consider the Wald test examining the difference in proportions between two arms:

$$Z = \frac{\hat{p}_E - \hat{p}_C}{\sqrt{\frac{\hat{p}_E \hat{q}_E}{n_E} + \frac{\hat{p}_C \hat{q}_C}{n_C}}}.$$

Under the null hypothesis, Z^2 is asymptotically chi-squared with one degree of freedom. Under the alternative hypothesis, power is an increasing function of the non-centrality parameter of the chi-squared distribution:

$$\begin{aligned}\text{Power } =& \frac{(p_E - p_C)^2}{p_E q_E / n_E + p_C q_C / nC} \\ =& \frac{n(p_E - p_C)^2}{\frac{p_E q_E}{\varphi + (n_E/n - \varphi)} + \frac{p_C q_C}{(1-\varphi) - (n_E/n - \varphi)}}.\end{aligned} \tag{1.11}$$

If we define a function

$$f(x) = \frac{(p_E - p_C)^2}{p_E q_E / (\varphi + x) + p_C q_C / [(1 - \varphi) - x]},$$

and the following Taylor's expansion:

$$f(x) = f(0) + f'(0)x + f''(0)x^2/2 + o(x^2),$$

then, after some calculations,

$$f'(0) = (p_E - p_C)^2 \frac{p_E q_E (1 - \varphi)^2 - p_C q_C \varphi^2)}{p_E q_E (1 - \varphi) + p_C q_C \varphi)^2},$$

$$f''(0) = -2(p_E - p_C)^2 \frac{p_E q_E p_C q_C}{((1-\varphi)\varphi)^3}.$$

Then, the non-centrality parameter is

$$\begin{aligned} n^{-1}\text{power } =& \frac{(p_E - p_C)^2}{p_E q_E/\varphi + p_C q_C/(1-\varphi)} \\ &+ (p_E - p_C)^2 \frac{p_E q_E (1-\varphi)^2 - p_C q_C \varphi^2}{p_E q_E(1-\varphi) + p_C q_C \varphi)^2}(n_E/n - \varphi) \\ &- (p_E - p_C)^2 \frac{p_E q_E p_C q_C}{((1-\varphi)\varphi)^3}(n_E/n - \varphi)^2 \\ &+ o((n_E/n - \varphi)^2) \\ =&(I) + (II) + (III) + o((n_E/n - \varphi)^2). \end{aligned} \tag{1.12}$$

The first term (I) represents the non-centrality parameter for a design targeting φ. The second term (II) represents the bias of the observed allocation from the optimal allocation. As the observed proportion deviates from target φ in either direction, the non-centrality parameter increases or decreases by term $(p_E - p_C)^2 \frac{p_E q_E (1-\varphi)^2 - p_C q_C \varphi^2}{p_E q_E(1-\varphi) + p_C q_C \varphi)^2}$. Note that this is only equal to 0 when we target Neyman allocation, with $\varphi = \frac{\sqrt{p_E q_E}}{\sqrt{p_E q_E} + \sqrt{p_C q_C}}$. Most procedures are asymptotically unbiased, so that $E(n_E/n - \varphi) = 0$. Then, the average power lost in a design is represented by (III), a direct function of the variability of the design. The impact of the variability of a randomization procedure on power will be demonstrated throughout this work, particularly in Chapters 2, 3, and 5.

1.3 Discussion

Adaptive designs allow trialists to achieve objectives such as maximizing power for a statistical test after observing patient responses, or minimizing the number of patients assigned to the inferior arm. However, their adoption has been slow in part because of FDA's historical resistance to adaptive designs. The continuing growth of research in this area has led to the FDA's current attitude, which is encouraging use of adaptive designs so long as Type I error can be controlled. Trialists tend to select simpler or more traditional trial designs because adaptive designs can result in various biases and statistical inference issues.

This chapter provided an overview of some common clinical trial designs. The designs covered included non-adaptive designs in Section 1.1 and some response-adaptive randomization designs in Section 1.2. The RAR designs target a specific allocation and assume responses between the experimental and control arms are independent. For example, RSIHR target allocation aims to minimize the total expected response of $\bar{Y}_E n_E + \bar{Y}_C n_C$ for two-arm trials with independent responses between the two groups.

Chapter 2 extends the RSIHR allocation scheme to the case that accounts for correlation between the two arms, which may arise due to a common exposure such as a common environmental setting (e.g. subjects of trial treated at the same hospital). Chapter 3 reviews desirability functions and their recent innovations and applications in medical research and public health; we show in Chapter 4 how they can be meaningfully used to effectively compare non-adaptive or adaptive designs across several characteristics and help the clinician arrive at a most appropriate design for his/her study. Chapter 5 contains a concrete application of desirability functions to design an improved study for an AIDS clinical trial.

Chapter 2

Optimal Allocation Proportions When Outcomes Between Treatment Arms Are Correlated

In this chapter, we derive a new target allocation that achieves the objective of RSIHR allocation (minimize total expected response for two-arm trials assuming smaller responses are better), yet additionally extends to the case that accounts for correlation between the two arms. Such a situation may arise due to a common exposure such as a shared environmental setting; for example, subjects in a trial are treated at the same hospital. A common assumption is that responses in the trial are independently and identically distributed (iid), which may be unrealistic in a clinical trial. This chapter examines how to allocate patients in two-arm clinical trials when the responses are continuous, smaller responses are better, and the iid assumption is violated.

The main results of this chapter demonstrate the importance of incorporating correlation in the treatment arms in the design of a clinical trial. In particular, we:

- first show it is feasible to minimize the total expected response $\bar{Y}_E n_E + \bar{Y}_C n_C$ when responses from the two arms are correlated by targeting a newly derived target allocation;
- second, targeting this optimal allocation proportion that adjusts for correlation results in smaller relative bias of the treatment effect;
- and third, failing to adjust the sample size when correlation is present can result in an underpowered study.

Section 2.1 provides an overview of correlation issues in two-arm trials. Section 2.2 considers trials where smaller responses are considered better, and derives the optimal allocation proportion that minimizes the total expected response when responses between the two arms are correlated. The various effects of increases in means and variances on this proportion are discussed in Section 2.3. Section 2.4 provides both parametric and nonparametric results of simulation studies that compare the newly derived optimal allocation proportion with a few of the other target allocations discussed in Chapter 1. Lastly, the chapter concludes with a Discussion.

2.1 Correlation in Two-Arm Trials

The Wald test is commonly used in clinical trials to test for a treatment effect between two treatment arms. A careful look at the Wald test highlights an important assumption that underlies the design and analysis of many clinical trials: outcomes of interest are independently and identically distributed (iid).

The focus of this chapter is to examine when the responses of two arms, Y_E and Y_C, are correlated due to some common exposure. Sources of common exposures could be environmental, social, economic, physical, etc. In clinical trials, correlation potentially arises due to common environmental exposures and nature of care (e.g. being treated at the same hospital), common drug exposures (e.g. concomitant meds), or common physical characteristics (e.g. shared genes in trials incorporating more than one family member). Large trials that are well-funded by industry tend to recruit from several diverse centers, with just a few patients per center. Consequently, correlation is not seen as a major concern in these cases, since these patients are in diverse settings and ought not to share many common exposures. However, in scenarios where trials are run on a smaller scale, such as in a single hospital system, the potential for common exposures in trial subjects should be properly accounted for in both the design and analysis stages.

Previous literature has argued that correlation between outcomes is acceptable when the randomization procedure employed ignores the correlation (Proschan) [69]. An alternative approach is to have a clustered randomized trial, where subjects in a cluster are considered exchangeable, but are independent across clusters. Common cluster units include location sites or similar baseline covariate values. The analysis that ignores clustering is acceptable if either there is no correlation within clusters, or no correlation between treatment assignment within clusters (Parzen) [65]. Otherwise, responses in a cluster are usually positively correlated, and a Wald test that ignores positive correlation is slightly conservative. The upshot is it is still common practice to have the independence assumption during statistical analysis, and sometimes even in planning.

However, it would be dismissive if we always settled for the independence assumption. In cases of negative correlation, the Wald test is anti-conservative and exhibits a weakness that can be improved upon should correlation be taken into account during the design stage. Furthermore, although previous work has shown that treating observations as independent is usually “acceptable” [69], Section 2.4 demonstrates that the bias of our treatment effect estimate can be significantly reduced should the targeted allocation proportion take into account the correlation between responses Y_E and Y_C from the E and C groups, respectively.

An example of correlation between treatment arms is highlighted by Biswas, et al. (2010) [14], using an example of a meta-analysis of 11 randomized trials investigating thrombolytic treatment versus placebo. Table 2.1 from their paper shows the number of all-cause mortality events out of the total number of patients for each treatment arm in each of the 11 studies:

	No. of Events / No of Participants	
Source	Treatment	Placebo
Mori et al., 1992	2/19	2/12
Haley et al., 1993	1/14	2/13
NINDS, 1995	76/312	87/312
ECASS, 1995	69/313	48/307
ECASS II, 1998	43/409	42/391
ATLANTIS A, 2000	16/71	5/71
ATLANTIS B, 1999	33/307	21/306
PROACT, 1999	7/26	6/14
PROACT II, 1999	29/121	16/59
Ancrod Stroke Study, 1994	8/64	14/68
STAT, 1999	83/248	82/252

Table 2.1: All-cause mortality associated with thrombolytic therapy in patients with acute ischemic stroke.

Given our usual assumption that responses from the patients in the two treatment arms are independent (since they are receiving different treatments), the estimated proportion of the number of events in the treatment arm and that in the placebo arm should be independent. However, the correlation coefficient of the responses in the two treatment arms is 0.666 [95 % CI: 0.149, 0.897, p-value = 0.01807], thus leading us to conclude that the responses of the patients in these two arms are significantly correlated [14]. Consequently, if a twelfth trial were to be run with similar enrollment criteria and treatments, we could fairly assume that the responses in the two groups of this twelfth trial are correlated.

Hanin discussed doubt of the validity of the iid assumption in the context of clinical trials, noting that trial participant selection is associated with disease characteristics and other medical conditions, which result in a source of dependence between individual response variables. Pre-randomization treatments provided to all of a trial’s participants, and the association between family history and the occurrence of health events, were also listed as reasons to question the independence assumption. In the case of certain educational and behavioural interventions, the intermingling of patients in different treatment arms post-randomization also introduces dependence [39]. In clinical trials studying infectious diseases, correlation may be present if participants have contact with a common infectious source, share tips on how to prevent infection, have spacial correlation, clustered data, or shared family structure [30]. Hanin claimed violation of the iid assumption as one of the reasons many clinical trial results are irreproducible [39].

How can we address correlation during the randomization process? Follman et al. (2014) provide an example of an NIH study that examined the effect of different factors on a measure of heart size. The study included members from 112 distinct families. Although this example was observational in practice, the authors discuss one way to incorporate familial correlation should this trial be a prospective trial with an intervention arm. The authors highlight a second example of studying efficacy of vaccines against malaria, and note that children of the same family or living in close proximity to each other tend to have similar attack rates. The authors discuss two randomization techniques that incorporate pairwise correlations and aim to minimize the variance of the treatment effect estimate. The first technique assumes all enrolling subjects' baseline covariates are available, and ranks permutations of pairwise correlations, randomizing in descending order of pairwise correlation rank one subject of each pair to one arm and the remaining subject to the other arm. Note then that these correlations are calculated from the covariances of baseline covariates, not the response variable. The second technique allows for sequential understanding of correlation, or ranks of correlations. An estimate of the pairwise correlation from the enrolled subjects is used to select or assign higher probability to a randomization assignment for the next subject that minimizes variance of the treatment effect estimate [30].

Biswas, et. al (2010) was the first to derive optimal allocation proportions for a two-arm clinical trial having correlated binary responses between the two arms [14]. Biswas, et al. (2011) discussed the non-correlated continuous outcome case [11]. Biswas, et al. (2005) had also derived solutions for three treatment arms and bivariate responses, where both components of the bivariate response were equally important [12], and responses between treatment arms were assumed independent. However, an examination of optimal allocation proportions for continuous responses with correlation between the two treatment arms has yet to be discussed in the literature. The following section derives the optimal allocation proportion that minimizes the total response $\bar{Y}_E n_E + \bar{Y}_C n_C$ when responses of the two arms are correlated and when smaller responses are considered better.

2.2 Derivation of the Optimal Allocation Proportion: Continuous Outcomes

We present the optimal allocation proportions for two-arm clinical trials with correlated, paired normal responses and evaluate the performance of various response-adaptive randomization (RAR) designs that target the optimal allocation proportions.

Consider a clinical trial with arms E and C, where there are n_E patients in experimental arm E and n_C patients in control arm C, so that $n_E + n_C = n$. Let $Y_E \sim N(\mu_E, \sigma_E^2)$ and $Y_C \sim N(\mu_C, \sigma_C^2)$ be continuous normal responses for groups E and C, respectively, and they are correlated with coefficient ρ. We can refer to historical data available prior to the clinical trial to estimate correlation ρ, pairing similar subjects by defining clear matching criterion such as tumor size, duration of infection, and family relationship.

We seek to derive the optimal allocation proportion, $R = \frac{n_E}{n_C}$, such that the total expected response is minimized. Since smaller responses are desirable, this means we want as many patients as possible to have better outcomes by choice of an optimal design. Thus, this objective may be viewed as incorporating an ethical consideration into the trial. Since we, without loss of generality, assume smaller responses are better, we seek to minimize $\mu_E n_E + \mu_C n_C$ subject to $Var(\hat{\psi}) = V\hat{a}r(\bar{Y}_E - \bar{Y}_C) = K$. This constraint is tantamount to pre-specifying the power of the test statistic. The derivation of the optimal allocation proportion proceeds as follows.

Set a fixed precision of the treatment effect, i.e. we set $n_E = \frac{Rn}{1+R}$ and $n_C = \frac{n}{1+R}$, and we have

$$\mu_E n_E + \mu_C n_C = \frac{\mu_E nR}{1+R} + \frac{\mu_C n}{1+R} = \frac{n}{1+R}(\mu_E R + \mu_C).$$

Let $K = Var(\hat{\psi}) = Var(\bar{Y}_E - \bar{Y}_C)$. Then $Var(\bar{Y}_E - \bar{Y}_C)$ is equal to

$$K = \frac{\sigma_E^2}{n_E} + \frac{\sigma_C^2}{n_C} - \frac{2\sigma_{EC}(1+R)}{\sqrt{n_E n_C}}.$$

$$nK = \frac{\sigma_E^2(1+R)}{R} + \sigma_C^2(1+R) - \frac{2\sigma_{EC}(1+R)}{\sqrt{R}}.$$

$$RK = R\left[\frac{\sigma_E^2(1+R)}{Rn} + \frac{\sigma_C^2(1+R)}{n} - \frac{2\sigma_{EC}\sqrt{1+R}}{\sqrt{Rn}}\frac{\sqrt{1+R}}{\sqrt{n}}\right].$$

$$RK = \frac{\sigma_E^2(1+R)}{n} + \frac{\sigma_C^2 R(1+R)}{n} - \frac{2\sigma_{EC}(1+R)\sqrt{R}}{n}.$$

Therefore, $\frac{n}{1+R} = \frac{\sigma_E^2}{RK} + \frac{\sigma_C^2 R}{RK} - \frac{2\sigma_{EC}\sqrt{R}}{RK}$, and

$$\frac{n}{1+R}(\mu_E R + \mu_C) = \left[\frac{\sigma_E^2}{RK} + \frac{\sigma_C^2 R}{RK} - \frac{2\sigma_{EC}\sqrt{R}}{RK}\right](\mu_E R + \mu_C).$$

Substituting $\frac{n}{1+R}$, we see that we must minimize

$$\frac{1}{K}\left[(\mu_E\sigma_E^2 + \mu_C\sigma_C^2) + \mu_E\sigma_C^2 R - 2\mu_E\sigma_{EC}\sqrt{R} - 2\mu_C\sigma_{EC}/\sqrt{R} + \frac{\mu_C\sigma_E^2}{R}\right] = f. \tag{2.1}$$

We minimize f with respect to R, set the derivative to zero, and multiply both sides by R^2 and see then that the optimal allocation $R = \frac{n_E}{n_C}$ is the solution to:

$$\frac{df}{dR}R^2 = \mu_E \sigma_C^2 R^2 - \mu_E \sigma_{EC} R^{3/2} + \mu_C \sigma_{EC}\sqrt{R} - \mu_C \sigma_E^2 = 0. \tag{2.2}$$

Equation 2.2 shows that the values of the means and variances of the responses in the experimental arm and the control arm, and the correlation between the responses of the two arms as can be seen by the covariance term σ_{EC}, will determine the value of the optimal allocation proportion R.

Equation 2.2 can be rewritten in terms of the ratios of the means and variances, $\mu^* = \frac{\mu_E}{\mu_C}$ and $\sigma^* = \frac{\sigma_E}{\sigma_C}$. To do so, divide both sides of the equation by μ_C to get

$$\frac{df}{dR}R^2 = \mu^* \sigma_C^2 R^2 - \mu^* \sigma_{EC} R^{3/2} + \sigma_{EC}\sqrt{R} - \sigma_E^2 = 0, \tag{2.3}$$

followed by division of both sides of the equation by σ_C^2 to get

$$\frac{df}{dR}R^2 = \mu^* R^2 - \mu^* \frac{\sigma_{EC}}{\sigma_C^2} R^{3/2} + \frac{\sigma_{EC}}{\sigma_C^2}\sqrt{R} - {\sigma^*}^2 = 0. \tag{2.4}$$

Because $\rho = \frac{\sigma_{EC}}{\sigma_E \sigma_C}$, the term $\frac{\sigma_{EC}}{\sigma_C^2} = \frac{\rho\sigma_E\sigma_C}{\sigma_C^2} = \frac{\rho\sigma_E}{\sigma_C} = \rho\sigma^*$. Equation 2.4 can then be rewritten as

$$\frac{df}{dR}R^2 = \mu^* R^2 - \mu^* \rho\sigma^* R^{3/2} + \rho\sigma^*\sqrt{R} - {\sigma^*}^2 = 0. \tag{2.5}$$

Equation 2.5 reveals that the optimal R is a function of the ratio of the means, the ratio of the variances, and the correlation. There are three analytical solutions of R that solve Equation 2.5; two are imaginary roots and only one is a real root. After very tedious algebra, the real solution for R is:

Corollary

$$\begin{aligned} R = &\frac{\rho^2 {\sigma^*}^2}{4} + \frac{1}{2}\sqrt{\Bigg(a + b/\big[\big(3(c+d)\big)^{1/3}\big] + \frac{1}{3 \times 2^{1/3}{\mu^*}^2}\big(c+d\big)^{1/3}} \\ &+ \frac{1}{2}\Bigg[2a - b/\big[\big(3(c+d)\big)^{1/3}\big] + \frac{1}{3 \times 2^{1/3}{\mu^*}^2}\big(c+d\big)^{1/3} \\ &+ \Bigg(\frac{8\rho^2{\sigma^*}^2}{{\mu^*}^2} - \frac{8\rho^2(-1+\rho^2){\sigma^*}^4}{\mu^*} + \rho^6{\sigma^*}^6\Bigg)\Bigg/4\sqrt{\big(a + b/\big[\big(3(c+d)\big)^{1/3}\big] + \frac{1}{3 \times 2^{1/3}{\mu^*}^2}\big(c+d\big)^{1/3}}\Bigg]^{1/2}, \end{aligned} \tag{2.6}$$

where

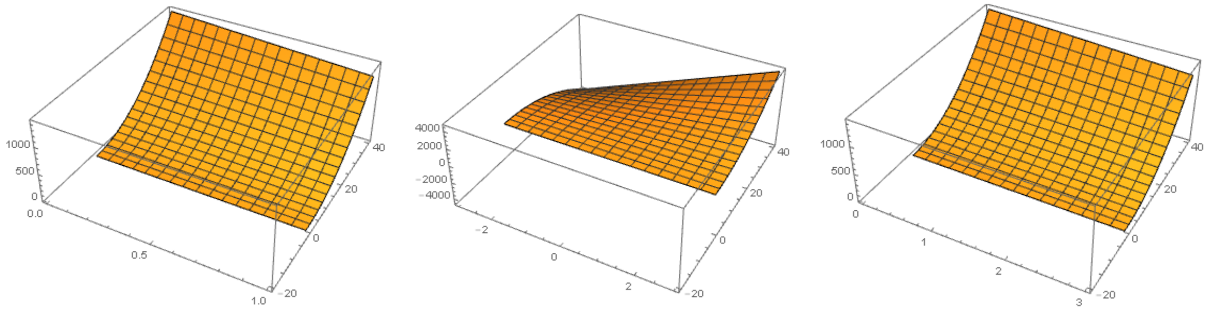
$$a = \frac{-4(-1+\rho^2){\sigma^*}^2}{3\mu^*} + \frac{\rho^4{\sigma^*}^4}{4};$$

$$b = \left(2^{(1/3)}(-4+\rho^2)^2{\sigma^*}^4\right);$$

$$c = 27{\mu^*}^2\rho^4{\sigma^*}^4 - 144{\mu^*}^3(-1+\rho^2){\sigma^*}^6 - 18{\mu^*}^3\rho^4(-1+\rho^2){\sigma^*}^6 + 16{\mu^*}^3(-1+\rho^2)^3{\sigma^*}^6 + 27{\mu^*}^4\rho^4{\sigma^*}^8;$$

$$d = \sqrt{-4(16\mu^{*2}\sigma^{*4} - 8\mu^{*2}\rho^2\sigma^{*4} + \mu^{*2}\rho^4\sigma^{*4})^3 + c^2}.$$

Figure 2.1 displays $\frac{df}{dR}R^2$ as a function of correlation and R, μ^* and R, and ${\sigma^*}^2$ and R. Figure 2.1a) displays correlation ranging between -1 and 1, b) shows μ^* ranging from -3 to 3, and c) exhibits ${\sigma^*}^2$ between 0 and 3. These plots indicate that R takes on positive values.



(a) correlation and R

(b) μ^* and R

(c) ${\sigma^*}^2$ and R

Figure 2.1: $\frac{df}{dR}R^2$ as a function of

It can be further shown that the real solution has interesting properties relating the behavior of R as correlation changes for relative relationships between the ratio of the means and the ratio of the variances. The solution for R satisfies the two behavior properties noted in the two bullet points below.

Corollary

- When $\frac{\sigma_C^2}{\sigma_E^2} > \frac{\mu_E}{\mu_C}$, R decreases as ρ increases.
- When $\frac{\mu_E}{\mu_C} = \frac{\sigma_C^2}{\sigma_E^2}$, $R = \frac{\sigma_E^2}{\sigma_C^2} = \frac{\mu_C}{\mu_E}$. R is then constant, regardless of ρ.

Table 2.2 shows how R changes when $\frac{\mu_E}{\mu_C} = 0.95$, with each column depicting a different value of $\frac{\sigma_E^2}{\sigma_C^2}$. This value of $\frac{\mu_E}{\mu_C}$ was chosen to represent a response in the experimental arm considered 5% better than that of the response in the control arm, since smaller values are considered better. Note that when $\frac{\mu_E}{\mu_C} = \frac{\sigma_C^2}{\sigma_E^2}$, $R = \frac{\sigma_E^2}{\sigma_C^2} = \frac{\mu_C}{\mu_E} = 1/0.95 = 1.053$, which is a constant, as expected, regardless of the value of ρ. The value of R depends on the ratio of the variances relative to the ratio of the means, and is further illustrated in Figure 2.3a. The solid red line marks the consistent value of R when $\frac{\sigma_E^2}{\sigma_C^2} = \frac{\mu_C}{\mu_E} = 1/0.95 = 1.053$. It can be seen that when $\frac{\sigma_C^2}{\sigma_E^2} > \frac{\mu_E}{\mu_C} = 0.95$, R decreases as ρ increases (the black, dark red, and navy curves below the solid red line). On the other hand, when $\frac{\sigma_C^2}{\sigma_E^2} < \frac{\mu_E}{\mu_C} = 0.95$, R increases as ρ increases (the green and orange curves above the solid red line). The smallest number of patients are placed in experimental group E when $\rho = 1$ and when $\frac{\sigma_E^2}{\sigma_C^2} = 0.5$. This means that when the responses of the two groups are perfectly correlated, if the responses in control arm C have twice the variance as the responses in the experimental arm E, then the control arm C needs twice the number of enrolled patients as in experimental arm E to minimize the total expected response from all patients.

	$R = \frac{n_E}{n_C}$ when $\frac{\sigma_E^2}{\sigma_C^2}$ equals					
ρ	0.5	0.75	1	1.053	1.2	1.5
-1.0	0.821	0.940	1.035	1.053	1.100	1.185
-0.9	0.814	0.936	1.034	1.053	1.101	1.189
-0.8	0.807	0.933	1.034	1.053	1.103	1.195
-0.7	0.799	0.928	1.033	1.053	1.105	1.200
-0.6	0.791	0.924	1.032	1.053	1.107	1.206
-0.5	0.782	0.919	1.031	1.053	1.109	1.213
-0.4	0.772	0.914	1.030	1.053	1.112	1.220
-0.3	0.762	0.908	1.029	1.053	1.114	1.228
-0.2	0.751	0.902	1.028	1.053	1.117	1.236
-0.1	0.739	0.896	1.027	1.053	1.120	1.246
0.0	0.725	0.888	1.026	1.053	1.124	1.257
0.1	0.711	0.881	1.025	1.053	1.128	1.268
0.2	0.696	0.872	1.023	1.053	1.132	1.282
0.3	0.679	0.862	1.021	1.053	1.137	1.297
0.4	0.660	0.852	1.019	1.053	1.142	1.314
0.5	0.640	0.840	1.017	1.053	1.149	1.333
0.6	0.617	0.826	1.015	1.053	1.156	1.356
0.7	0.592	0.811	1.012	1.053	1.164	1.383
0.8	0.565	0.793	1.009	1.053	1.174	1.414
0.9	0.534	0.773	1.005	1.053	1.186	1.453
1.0	0.500	0.750	1.000	1.053	1.200	1.500

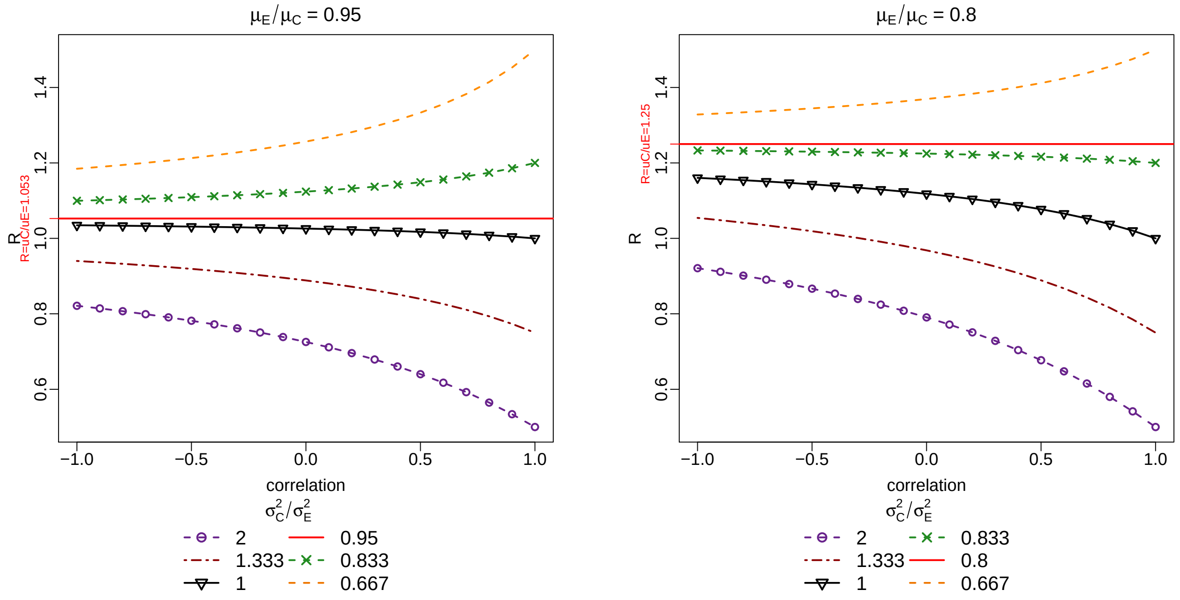
Table 2.2: Values of the optimal ratio $R = \frac{n_E}{n_C}$ for $\frac{\mu_E}{\mu_C} = 0.95$ and varying $\frac{\sigma_E^2}{\sigma_C^2}$.

Table 2.2 and Figure 2.3a show how R varies with different ratios of variances when the responses in group E are 5% better than those in group C. Table 2.3 and Figure 2.3b display values of R for a more extreme difference in the responses from the two groups, with $\mu_E/\mu_C = 0.80$. When $\sigma_E^2/\sigma_C^2 = \mu_C/\mu_E = 1/0.8$, $R = 1/0.8 = 1.25$, regardless of correlation. This is depicted by the solid red line in Figure 2.3b. It can also be seen that when $\sigma_C^2/\sigma_E^2 > \mu_E/\mu_C = 0.80$, R increases as correlation increases (the orange curve above the solid red line). On the other hand, when $\sigma_C^2/\sigma_E^2 < \mu_E/\mu_C = 0.80$, R decreases as correlation increases (the green, black, dark red, and navy curves below the solid red line).

The next section discusses how R changes with changes in the ratio the of means μ_E/μ_C and with changes in the ratio of the variances σ_C^2/σ_E^2.

	$R = \frac{n_E}{n_C}$ when $\frac{\sigma_E^2}{\sigma_C^2}$ equals					
ρ	0.5	0.75	1	1.2	1.25	1.5
-1.0	0.921	1.054	1.160	1.233	1.250	1.328
-0.9	0.911	1.048	1.157	1.233	1.250	1.331
-0.8	0.901	1.042	1.154	1.232	1.250	1.334
-0.7	0.890	1.035	1.151	1.231	1.250	1.337
-0.6	0.879	1.027	1.147	1.231	1.250	1.341
-0.5	0.867	1.019	1.143	1.230	1.250	1.345
-0.4	0.854	1.010	1.139	1.229	1.250	1.349
-0.3	0.840	1.001	1.134	1.228	1.250	1.353
-0.2	0.824	0.991	1.129	1.227	1.250	1.358
-0.1	0.808	0.980	1.124	1.226	1.250	1.363
0.0	0.791	0.968	1.118	1.225	1.250	1.369
0.1	0.772	0.955	1.111	1.223	1.250	1.376
0.2	0.751	0.941	1.104	1.222	1.250	1.383
0.3	0.728	0.925	1.096	1.220	1.250	1.392
0.4	0.704	0.908	1.087	1.219	1.250	1.401
0.5	0.677	0.889	1.077	1.216	1.250	1.412
0.6	0.648	0.867	1.066	1.214	1.250	1.424
0.7	0.615	0.843	1.053	1.211	1.250	1.438
0.8	0.580	0.816	1.038	1.208	1.250	1.455
0.9	0.541	0.785	1.020	1.204	1.250	1.475
1.0	0.500	0.750	1.000	1.200	1.250	1.500

Table 2.3: Optimal values of the ratio $R = \frac{n_E}{n_C}$ for $\frac{\mu_E}{\mu_C} = 0.80$ and varying $\frac{\sigma_E^2}{\sigma_C^2}$.



(a) $\mu_E/\mu_C = 0.95$.

(b) $\mu_E/\mu_C = 0.80$.

Figure 2.3: R for $\frac{\mu_E}{\mu_C} = 0.95$ and $\frac{\mu_E}{\mu_C} = 0.80$ and varying $\frac{\sigma_C^2}{\sigma_E^2}$.

2.3 Changes in R Due to Changes in Means or Variances

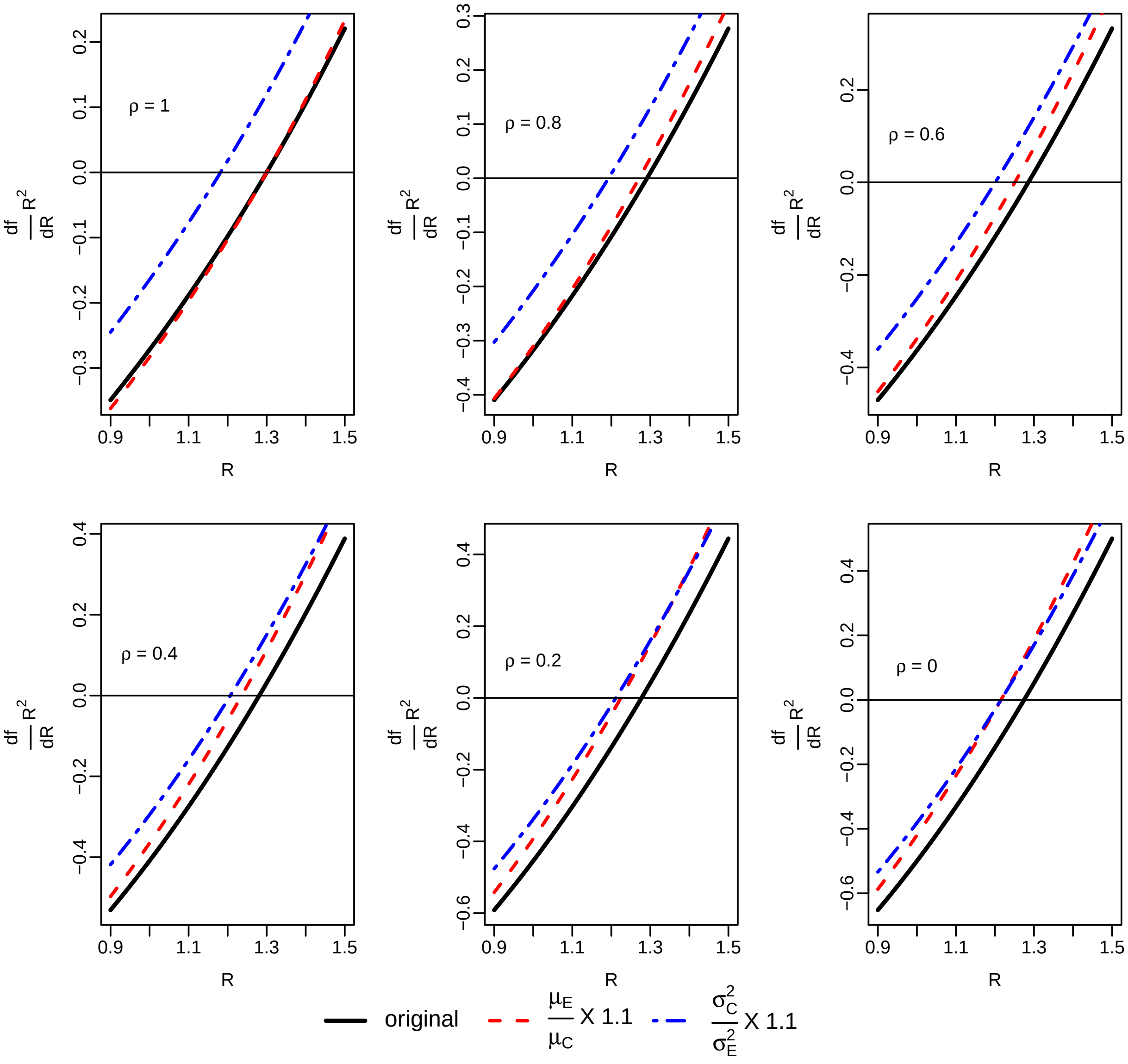
Section 2.2 shows analytically that R depends on the ratio of the means, the ratio of the variances, and the correlation ρ. This section studies the changes in R associated with a change in ratio of means, and with a change in ratio of variances. For example, if $\mu_E/\mu_C = 0.80$ (20% less expected response in the experimental arm), and $\sigma_E^2/\sigma_C^2 = 1.3$ (30% larger variance in experimental arm), how does R change with an increase in the ratio of means by 10%, so that $\mu_E/\mu_C = 0.8 \times 1.1 = 0.88$? From the Lemma in Section 2.2, we see that the ratio of μ_E/μ_C is compared with the ratio of σ_C^2/σ_E^2. How does R change when *this* ratio increases by 10%, so that $\frac{\sigma_C^2}{\sigma_E^2} = \frac{1}{1.3} \times 1.1 = 0.846$? Which change in R is more significant?

To assess the effect of a change in ratio of means on R, μ_C is set to equal 1, without loss of generality, while μ_E is replaced with $\mu_E * (1 + \text{percent change})$ in Equation 2.2. The solution is termed $R1$. Similarly, to assess the effect of a change in ratio of variances on R, σ_C^2 is set to equal 1, without loss of generality, while σ_E^2 in Equation 2.2 is replaced with $\sigma_E^2/(1 + \text{percent change})$. The solution is termed $R2$.

We conclude that when $\rho > 0$, the effect of a percent increase in $\frac{\sigma_C^2}{\sigma_E^2}$ on R is larger than the the effect of the same percent increase in $\frac{\mu_E}{\mu_C}$ on R. In other words, when $\rho > 0$, $R2/R > R1/R$. On the other hand, when $\rho < 0$, a percent increase in the ratio of means $\frac{\mu_E}{\mu_C}$ has a larger impact on R than the same percent increase in the ratio of variances $\frac{\sigma_C^2}{\sigma_E^2}$. This means that when $\rho < 0$, $R1/R > R2/R$. When $\rho = 0$, the same percent increase in ratio of means and ratio of variances results in the same change in R, such that $R1/R = R2/R$.

An example for the behavior of changes in R can be seen in Figure 2.4, which illustrates with the solid black line the function of Equation 2.2 for original values of $\mu_E/\mu_C = 0.8$, $\sigma_E^2/\sigma_C^2 = 0.8$ for varying levels of positive correlation. The red dashed line shows the same function when μ_E/μ_C increases by 10%. We can see that as correlation increases, a 10% increase in ratio of means leads to decreased solutions for R. The blue two-dash line represents the same function for an equivalent increase of 10% on the ratio of σ_C^2/σ_E^2. As correlation increases, R decreases, placing fewer and fewer subjects in the experimental arm. Note that for positive correlations, the blue two-dash line is further from the original black solid line than is the red dashed line, indicating the higher impact of the same percent change of the ratio of variances on R. When the correlation is zero, the solution for R is the same for the same increase in ratio of means and in ratio of variances.

The values for R, $R1$, and $R2$ that solve Equation 2.2 (where $\frac{\partial f}{\partial R} R^2 = 0$ in Figure 2.4 and where$f$ is defined by Equation 2.1) are shown in Table 2.4. For example, when $\rho = 0.4$, $\mu_E/\mu_C = 0.8$, and $\sigma_E^2/\sigma_C^2 = 1.3$, $R = 1.2810$. If μ_E/μ_C is increased by 10%, the optimal allocation becomes $R1 = 1.2246$, placing fewer subjects in the experimental arm. The change in R is even larger if σ_C^2/σ_E^2 increases by 10%, with the optimal allocation dropping to $R2 = 1.2117$.



.

Figure 2.4: The function $\frac{\partial f}{\partial R}R^2$ for varying levels of correlation ρ when $\frac{\mu_E}{\mu_C} = 0.8$, $\frac{\sigma_E^2}{\sigma_C^2} = 1.3$, (1 + percent change in ratios) = 1.1. The value of R when $\frac{\partial f}{\partial R}R^2 = 0$ is the solution for R in Equation 2.2.

ρ	1	0.8	0.6	0.4	0.2	0
R	1.3000	1.2916	1.2855	1.2810	1.2775	1.2748
R1	1.3000	1.2712	1.2510	1.2361	1.2246	1.2154
R2	1.1818	1.1929	1.2009	1.2069	1.2117	1.2154
R1/R	1.0000	0.9842	0.9732	0.9649	0.9585	0.9535
R2/R	0.9091	0.9236	0.9342	0.9422	0.9484	0.9535

Table 2.4: Solutions for R, $R1$, and $R2$ for $\frac{\mu_E}{\mu_C} = 0.8$, $\frac{\sigma_E^2}{\sigma_C^2} = 1.3$, (1 + percent change) = 1.1, as depicted in Figure 2.4.

The larger impact on $R2$ than on $R1$ as shown in Figure 2.4 are similar regardless of the value of when μ_E/μ_C, $\sigma_E^2/\sigma_C^2 < 1$, and whether the percent change is positive or negative.

2.4 Performance of Optimal Allocation Proportion Adjusting for Correlation Relative to Other Common Target Allocations

There are two ways to utilize the new optimal allocation proportion R that adjusts for correlation, and found from solving Equation 2.2 (the solution is hereby termed $R.corr$). The first is to identify the sample size n needed to achieve a given power and maintain Type I error at a pre-specified level, and then to place $\frac{Rn}{1+R}$ in the experimental arm E, and $\frac{n}{1+R}$ in the control arm C. The second is to target R.corr using a Response-Adaptive Randomization Procedure (RAR), similar to targeting Neyman and RSHIR optimal proportions by RAR designs DBCD or ERADE in Section 1.2.

However, the target allocations reviewed in Section 1.2 did not adjust for or require a value for correlation; they depended on estimates for means and variances. To target R.corr, in the context of this dissertation, correlation is held constant and assumed known throughout the study. In practice, correlation can be estimated from prior studies or literature, or sequentially estimated within interim analyses of a multi-stage study.

In order to compare the performance of the optimal allocation proportion adjusting for correlation with other common target allocations, we examine the total expected response for each allocation under the null and alternative hypotheses. The null hypothesis is equality of the responses from the two rams. Since R.corr is derived with the objective of minimizing $\bar{Y}_E n_E + \bar{Y}_C n_C$ when correlation is present, it ought to have the lowest total response in our comparison study. In addition to examining whether R.corr achieves this, we further define other performance metrics not related to the optimization problem yet still important to overall assessment of the design. In these performance metrics, β_1 is the value of the true treatment effect, and $\hat{\beta_1}$ is the estimated treatment effect. The performance metrics are quantified by:

- bias under $H_0 = E(\hat{\beta}_1 - \beta_{1_{H_0}})$,

- relative bias under $H_1 = \mathrm{E}(\frac{\hat{\beta_1} - \beta_{1_{H_1}}}{\beta_{1_{H_1}}}) \times 100$,
- Type I error under H_0: proportion of times the null hypothesis was falsely rejected in 10,000 simulated clinical trials of the same scenario,
- power under H_1: proportion of times the null hypothesis was correctly rejected over 10,000 simulated clinical trials of the same scenario.

However, the Type I error and power are both calculated from the T test statistic. Is this valid given the correlation structure simulated in the data? The violation of the independence assumption renders many parametric analyses (standard Gaussian, Student's t, Chi-squared) inadequate, since they depend at their core on the independence postulate. Biswas (2010) did not discuss this in his derivation of optimal allocation proportion in trials with correlated binary outcomes [14], and we address this issue here. How can we assess the impact of the misspecification of the test statistic distribution on the outcome of our analyses? Standard assumptions regarding the test statistic distribution lead to commonly reported results: p-value under the null, power under the alternative, sample size, and estimates of and confidence intervals for parameters of interest (in this work, namely, the treatment effect). Hanin (2017) notes that neither the correlation coefficient for a pair of individual observations, nor the distance between their distributions (ϵ), is estimable from the observations alone. The author goes on to note that - hypothetically - even if we did know the correlation and underlying distributions of a pair of observations, we would believe our statistical analyses to be robust if the distribution, P_0, of the test statistic under the iid assumption is close to the "true" distribution of the same hypothesis *without* the iid assumption. However, obtaining estimates of the deviation of the output of the statistical analysis (e.g. some test statistic) as a function of ϵ "in most cases goes far beyond the reach of contemporary probability theory and statistics" [39].

In this work, the T test statistic is the primary source used in analyses of a treatment effect. Because deriving the theoretical deviation of the T test statistic as a function of the ϵ is difficult, we can still look at different scenarios and compare the distribution of the T test statistic under the iid assumption (P_0) with that of the T test statistic in the presence of correlation. To do so, we study the T test statistic distributions under three scenarios:

1. $Y_E \sim N(\mu_E, \sigma_E^2)$, $Y_C \sim N(\mu_C, \sigma_C^2)$; $Y_E \perp\!\!\!\perp Y_C$ (independent); α is a pre-specified Type I error rate, $\tilde{\beta}$ is a pre-specified Type II Error rate, with $1 - \tilde{\beta}$ equaling power. Total Sample Size for a balanced trial derived from the formula:

$$n = n_E + n_C = 2\frac{(z_{1-\alpha/2} + z_{1-\tilde{\beta}})^2(\sigma_E^2 + \sigma_C^2)}{(\mu_E - \mu_C)^2}. \tag{2.7}$$

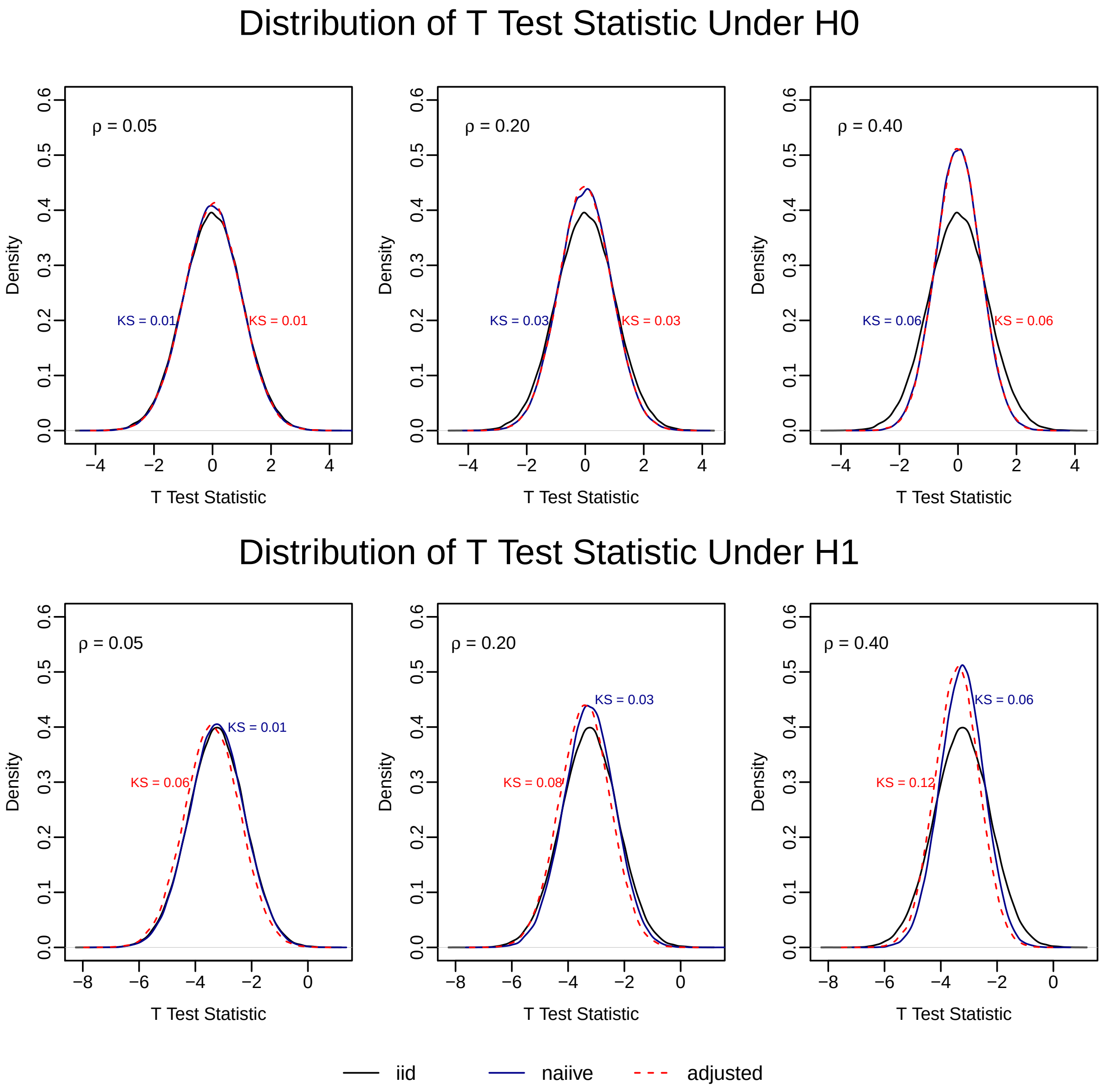
Half are allocated to treatment arm E, the other half to treatment arm C. The distribution of the T test statistic is termed P_0.

2. $Y_E \sim N(\mu_E, \sigma_E^2)$, $Y_C \sim N(\mu_C, \sigma_C^2)$; $\rho_{EC} \neq 0$; Total Sample Size based on Equation B.1, with half allocated to treatment arm E, the other half to treatment arm C.

3. $Y_E \sim N(\mu_E, \sigma_E^2)$, $Y_C \sim N(\mu_C, \sigma_C^2)$; $\rho_{EC} \neq 0$; Total Sample Size derived from Equation B.1, but the allocation proportion is R.corr, optimized for correlated outcomes based on Equation 2.2.

In the study of a case with a treatment effect of -5, where $\mu_E = 127, \mu_C = 132, \sigma_E^2 = 330, \sigma_C^2 = 235$, scenario #2 and #3 under the null hypothesis are equivalent (with identical Kolmogorov-Smirnov distance, as shown at the top of Figure 2.5), and under the alternative hypothesis they yield very similar distributions. In scenario #3 we observe a higher peak which borrows slightly from the tails. Figure 2.5 displays distributions of the T Test Statistic in the iid, naive, and adjusted analyses, and shows that under the alternative hypothesis, the test statistic accounting for correlation shifts towards the direction of the alternative. The Kolmogorov-Smirnov distance from P_0 (the distribution depicted by the black solid curve) increases from 0.06 to 0.12 as correlation increases from 0.05 to 0.40, thus indicating that the distribution of the T test statistic shifts further and further away from P_0 as correlation increases. The distribution of the test statistic in scenario #3 is closer to P_0 on the left tail when compared to the distance from P_0 on the right tail. This indicates that working under the iid framework is less of a deviation from one that adjusts for the correlation as the data supports or moves in the direction of the alternative.

Similarly, Figure 2.6 depicts the distribution of $\hat{\beta_1}$ under the null and alternative hypotheses. The Kolmogorov-Smirnov distance reveals that the distribution of scenario #3 deviates further from that of the independent case P_0 than does scenario #2. However, the figure allows us to make two decisions: first, it is a safe assumption that the distribution of the coefficient estimate is still normal; second, the deviation from the iid case of scenario #1 is not too alarming, especially relative to the density curves that show the distributions of the estimate of log odds for various levels of correlation as shown in Biswas (2010) [[14]]. Given this observation, we will proceed in Section 2.4.1 with traditional parametric analysis when comparing the performance of the optimal allocation proportion that accounts for correlation with other target allocations that may not. Section 2.4.2 compares the performance using bootstrapped confidence interval widths.

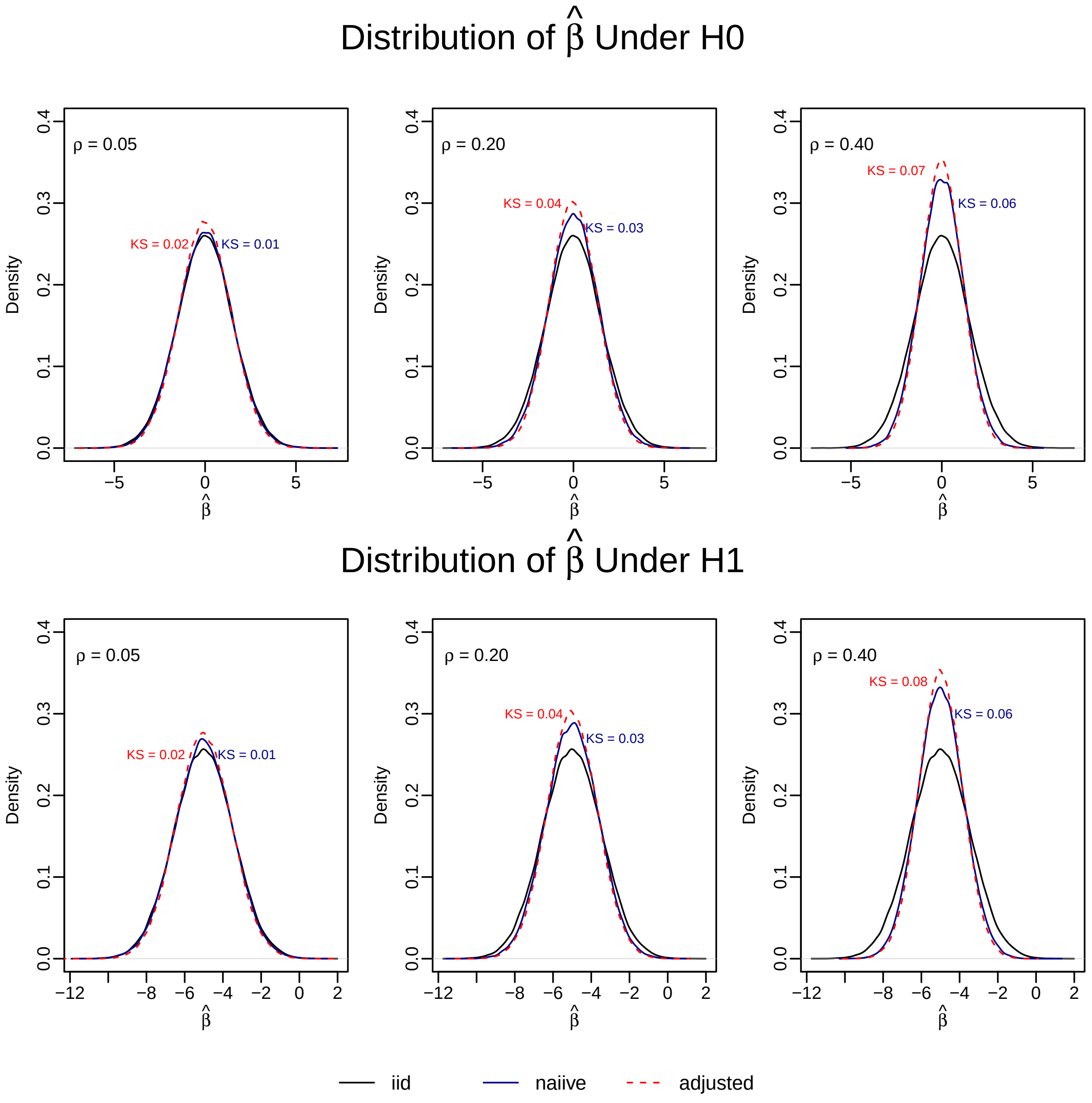


Distribution of T Test Statistic Under H0

Distribution of T Test Statistic Under H1

.

Figure 2.5: Distributions of the T Test Statistic under $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = -5$ for $\rho \in (0.05, 0.20, 0.40)$.



.

Figure 2.6: Distributions of $\hat{\beta_1}$ under $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = -5$ for $\rho \in (0.05, 0.20, 0.40)$.

2.4.1 Parametric Methods in Assessing of Optimal Allocation Proportion Accounting For Correlation

We now assess the performance of a design targeting R.corr relative to that of designs that assume outcomes are independent. The comparative designs considered are Complete Randomized Design (CRD), Permuted Block Design (PBD) with a block size of 8 (see Section 1.1), and Doubly Biased Coin Design (DBCD) targeting allocations that do not account for correlation: Neyman, RSIHR, and RSIHR2 allocations (see Section 1.2).

To compare the designs, we evaluate the total expected response under the six aforementioned clinical trial designs in two cases with different levels of correlation. Let

$$\boldsymbol{X} = \begin{bmatrix} 1 & T_1 \\ 1 & T_2 \\ \vdots & \vdots \\ 1 & T_j \\ \vdots & \vdots \\ 1 & T_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},$$

where $T_j = \mathbb{1}(\text{ subject } j \text{ is in experimental arm E})$. Then, the response Y is modeled by:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Let $\theta^T = (\mu_E, \mu_C, \sigma_E^2, \sigma_C^2)$. The below cases are examined and motivated from two real trials [56, 68]:

- Contraceptive Study: $\theta^T = (127, 132, 330, 235), n = 355$.

  $H_0 : \beta_1 = 0, H_1 : \beta_1 = -5$.

  [a.] $\rho = 0.05$.

  [b.] $\rho = 0.30$.

- Scleroderma Study: $\theta^T = (21.8, 27.5, 219, 144), n = 165,$.

  $H_0 : \beta_1 = 0, H_1 : \beta_1 = -5.7$.

  [a.] $\rho = 0.05$.

  [b.] $\rho = 0.30$.

  [c.] $\rho = 0.60$.

The data for the contraceptive study come from a study evaluating the association between oral contraceptive pills and blood pressure [56], and the data for the scleroderma study come from a clinical trial looking at Methotrexate versus placebo for scleroderma patients [68]. In both cases, we remain consistent with calling the experimental arm group E, and the control arm group C. The two cases show that the higher the correlation between responses in the two arms, the more the design targeting R.corr can reduce the total response under the alternative hypothesis.

The sample sizes are pre-determined assuming responses are independent with the given means and variances of the two groups under the alternative hypothesis, with Type I error controlled at $\alpha = 0.05$ and power at 80%. Simulations are then based on $n = 355$ subjects for the contraceptive study and $n = 165$ subjects for the scleroderma study. Instead of allocating half the subjects in each arm, we seek to minimize the total expected response of the two groups using the optimal allocation proportion R.corr.

The optimal allocation proportion accounting for correlation, termed R.corr, is shown for varying levels of correlation in Table 2.5 for the contraceptive study, and in Table 2.11 for the scleroderma study. In the contraceptive study, the R ranges from 1.2081 when correlation is 0, to 1.4042 when correlation is 1. When $\rho = 0.05$, $R = 1.2128$, while when $\rho = 0.30$, $R = 1.2407$.

ρ	R.corr
0.00	1.2081
0.05	1.2128
0.10	1.2177
0.15	1.2230
0.20	1.2285
0.25	1.2344
0.30	1.2407
0.35	1.2473
0.40	1.2545
0.45	1.2621
0.50	1.2703
0.55	1.2792
0.60	1.2887
0.65	1.2990
0.70	1.3102
0.75	1.3224
0.80	1.3358
0.85	1.3504
0.90	1.3665
0.95	1.3844
1.00	1.4042

Table 2.5: Contraceptive study: optimal allocation proportion for varying levels of correlation with $\theta^T = (127, 132, 330, 235)$ and $n = 355$.

In each evaluation, the outcomes are simulated to be correlated with ρ as specified in each case. Each simulation for a particular design includes 10,000 iterations - each iteration representing a single clinical trial. In order to introduce correlation into the simulated data, a set of outcomes of length n is simulated for *each* treatment group, where the outcome is simulated as such:

- Independently generate a vector $Z_1 \sim N(0,1)$ of length n and a second vector $Z_2 \sim N(0,1)$ also of length n.
- Let $Z_3 = \rho Z_1 + \sqrt{1-\rho^2} Z_2$.

- Then n outcomes for arm E are calculated as $Y_E = \mu_E + \sqrt{\sigma_1^2} Z_1$, while the n outcomes for arm C are calculated as $Y_C = \mu_C + \sqrt{\sigma_2^2} Z_3$.

The Doubly Biased Coin Design (DBCD) is simulated as discussed in Section 1.2.2, with $m_0 = 5$ and, $\gamma = 2$. The value of m_0 is specified by the trialist; one might consider to have larger values of m_0 if the responses between subjects are considered highly variable, so that the estimates incorporated into R.corr are more reliable and so that the target allocation proportion at the start of the response-adaptive portion of the trial is less variable. After the first $2m_0 = 10$ patients are enrolled, R of R.corr is derived by solving Equation 2.2 for each enrolling patient using estimates of the means and variances of the previously enrolled patients. In this work, correlation is assumed to be known and held constant, although iteratively estimating correlation with sequential looks at the data can be done in practice. Two assumptions are taken in these simulations: first that the response of patient $j-1$ can be observed immediately and prior to the enrollment of patient j, and that we can assess the data immediately after an observed response. In practice, the timing of interim looks ought to be pre-defined, and patient j may be allocated using data from earlier patients 1 through $j-k$, where $k \geq 1$.

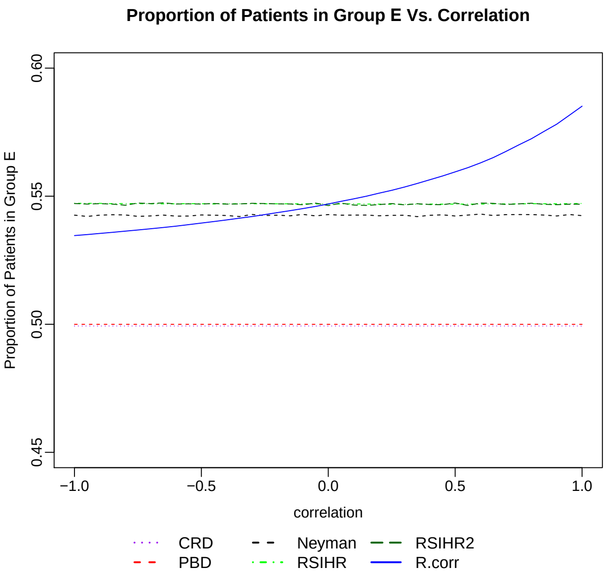
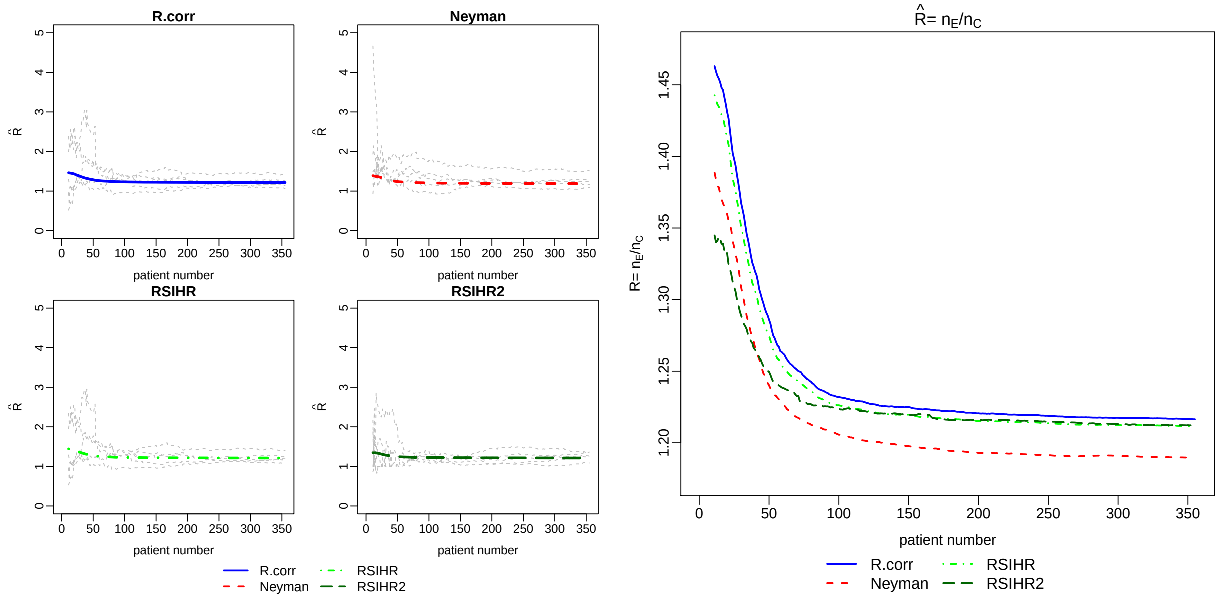


Figure 2.7: Contraceptive study: proportion of patients in the experimental arm E for different levels of correlation, averaged across 10,000 iterations.

After successfully enrolling a total of n patients, Figure 2.6 shows the proportion of patients enrolled in each treatment arm and the total expected responses. The straight horizontal lines shown in Figure 2.7 reassure us that in our simulations, the proportion of patients in the experimental arm for designs that do not adjust for correlation were consistent even as we have incorporated correlation into the simulations.

The blue solid curve depicting proportion of patients in group E under R.corr shows increasing patients in group E as correlation increases. Note for negative correlation values, R.corr allocates fewer subjects to the experimental arm than does the other target allocations Neyman, RSIHR, and RSIHR2. Since negative correlation is not common in health outcomes, we focus on $\rho \geq 0$ in this work.



(a)

(b)

Figure 2.8: Contraceptive study: (a) $\hat{R} = n_E/n_C$ across 355 enrolled patients, $\rho = 0.05$. Each dotted gray line represents $\hat{R}$ in a single iteration. The bolder lines in color represent the average $\hat{R}$ for the design across 10,000 iterations.
(b) $\hat{R} = n_E/n_C$ across 355 enrolled patients averaged across 10,000 iterations.

The value of $\hat{R}(j) = \frac{n_E(j)}{n_C(j)}$ for each patient j in the contraceptive study with $\rho = 0.05$ is plotted in Figure 2.8a, where each dotted gray line represents the results of a single simulated study or iteration, and the bolder lines in color represent the average $R(j) = \frac{n_E(j)}{n_C(j)}$ for each patient j across 10000 iterations. Figure 2.8b summarizes the average $\hat{R} = \frac{n_E(j)}{n_C(j)}$. We can see that overall, R.corr seeks to place more subjects in the experimental arm. In this scenario, the gap between R widens between R.corr and Neyman allocation, and narrows between R.corr and RSIHR2. The narrowing gap between R.corr and RSIHR2 is not surprising since RSIHR2 seeks to minimize the expected total response and to place fewer patients in the inferior arm. As more data becomes available with more enrolled subjects, evidence points towards the control arm being inferior, pushing RSIHR's R to place more patients in the experimental arm, thus approaching the R resulting from R.corr. With R.corr consistently having higher R than other designs, the probability of being assigned to the experimental arm is always higher for each patient in the R.corr design than in the others .

Table 2.6 shows the total expected responses of the simulated contraceptive study under the alternative

	Patients in E (H1)			**Responses**		**Total Response**
	mean	sd	proportion	μ_E	μ_C	$\mu_E n_E + \mu_C n_C$
$\rho = 0.05$						
CRD	177.256	9.392	0.499	127.007	132.021	45978.777
PBD	177.487	0.810	0.500	127.006	132.023	45977.609
Neyman	192.598	8.540	0.543	127.009	132.000	45898.678
RSIHR	194.178	8.418	0.547	127.026	132.023	45897.802
RSIHR2	194.294	8.429	0.547	127.011	132.011	45892.506
R.corr	194.535	8.606	0.548	127.024	132.022	45895.534
$\rho = 0.30$						
CRD	177.256	9.392	0.499	127.007	132.020	45978.470
PBD	177.487	0.810	0.500	127.006	132.021	45977.342
Neyman	192.493	8.493	0.542	126.993	132.008	45897.362
RSIHR	194.127	8.464	0.547	127.028	132.019	45897.985
RSIHR2	194.051	8.308	0.547	127.037	131.989	45894.980
R.corr	196.516	9.671	0.554	127.029	132.021	45886.340

Table 2.6: Contraceptive study: total observed response under the alternative hypothesis $H_1 : \beta_1 = -5$; $\theta^T = (127, 132, 330, 235)$ and $n = 355$.

hypothesis. The comparison with RSIHR allocation is emphasized, as its goal is to minimize the total expected response, yet does not account for correlation. When $\rho = 0.05$, accounting for correlation using R.corr results in lower total response of 45895.534 relative to the total response of 45897.802 under the RSIHR design which assumes independence. The lowest total response of 45892.506 results from RSIHR2 allocation. Although R.corr on average placed more patients in group E, RSIHR2 yielded a smaller response due to its μ_E and μ_C values of 127.011 and 132.011, respectively, both smaller than the average responses of R.corr. Excluding RSIHR2 allocation, R.corr target does succeed in having smaller total responses than the other designs evaluated.

When correlation increases to $\rho = 0.30$ under the alternative hypothesis, R.corr has the lowest response of all designs evaluated, with the total response being 45886.340, compared with 45897.985 from RSIHR allocation. The differences between the expected total responses is larger, highlighting the value of using R.corr especially for larger correlation values. Note the large improvement of decreasing the expected total response of R.corr relative to non RAR designs CRD and PBD, which had expected total responses of 45978.470 and 45977.342, respectively. One may consider the difference in total response between RSIHR and R.corr to be quite small, which can be attributed to the small effect size, but we will see a larger difference later in the scleroderma case when the effect size is moderate.

In addition to examining performance of R.corr with respect to the objective of minimizing the expected total response, we should consider other statistics as well. A Wald test is performed to test for a treatment effect between the two treatment arms. The results shown in Tables 2.7 and 2.8 include the average treatment effect estimate ($\hat{\beta_1}$), the average standard error of the treatment effect estimate ($E(se(\hat{\beta_1}))$), the standard deviation of the treatment effect estimate (sd) across 10,000 iterations, the ratio of the standard deviation to standard error (sd_se), bias, relative bias (rb), mean squared error (mse) of the treatment effect, power, and empirical coverage (ci). The definitions of these measures and metrics in the context of this work are:

- *Standard Deviation to Standard Error ratio* (sd_se) $= \frac{SD(\hat{\beta_1})}{E(SE(\hat{\beta_1}))}$. A ratio close to 1 indicates no noticeable biases and reasonable estimated standard errors during the simulation. When this ratio is close to 1, we expect well-controlled empirical test sizes and coverage. A ratio greater than 1 indicates underestimation of standard errors, resulting in increased Type I error and lower empirical coverage. When the ratio is less than 1, the standard errors have been overestimated, resulting in lower empirical Type I errors and higher empirical coverage.

- *Bias* (bias) $= E(\hat{\beta_1} - \beta_1)$ is calculated as the estimated treatment effect less the true treatment effect value.

- *Relative bias* (rb) $= E(\frac{\hat{\beta_1}-\beta_1}{\beta_1}) \times 100$ is calculated as the (true treatment effect - estimated treatment effect)/(true treatment effect) × 100. Since the true treatment effect under the null hypothesis is zero, the relative bias term is only defined under the alternative hypothesis.

- The *mean squared error* (mse) $= E(\hat{\beta_1} - \beta_1)^2 + Var(\hat{\beta_1})$ is mean squared error of the treatment effect, calculated as the squared bias term plus the variance of the treatment effect estimate. This is a meaningful metric since for unbiased estimators, the mean squared error approaches $Var(\hat{\beta_1})$ as $n \to \infty$. Note that systematically large SE estimates are indicative of a lack of relative efficiency and would yield larger MSEs.

- *Size* (size) is the empirical test size, reporting Type I error under the null hypothesis, and power under the alternative hypothesis. This is the proportion of trials in which the null hypothesis is rejected during 10,000 simulated trials of a particular design.

- *Empirical coverage* (ci) is the percentage of time the (1-α)% confidence interval includes the true value of the treatment effect.

Table 2.7 shows results for the contraceptive study when $\rho = 0.05$. The first observation is that the sample size used (n=355) attains Type I error of 0.05 and power of 0.80 under CRD, as intended. The RAR designs targeting Neyman, RSIHR, RSIHR2, and R.corr all have lower power than 0.80. This is due to a loss in power in RAR designs caused by the variability of $R = \frac{n_E}{n_C}$ (see Section 1.2.3) [55]. PBD consistently

	$\hat{\beta_1}$	se($\hat{\beta_1}$)	sd	sd_se	bias	rb	mse	size	ci
Under H_0									
CRD	-0.014	1.785	1.777	0.996	-0.014		6.348	0.050	95
PBD	-0.017	1.783	1.773	0.994	-0.017		6.327	0.049	95
Neyman	-0.001	1.804	1.774	0.984	-0.001		6.407	0.049	95
RSIHR	-0.008	1.804	1.766	0.979	-0.008		6.379	0.045	95
RSIHR2	0.008	1.794	1.782	0.993	0.008		6.397	0.047	95
R.corr	-0.008	1.805	1.767	0.979	-0.008		6.384	0.046	95
Under H_1									
CRD	-5.014	1.785	1.777	0.996	-0.014	0.289	6.348	0.800	95
PBD	-5.017	1.783	1.773	0.994	-0.017	0.340	6.327	0.800	95
Neyman	-4.991	1.804	1.772	0.982	0.009	-0.177	6.399	0.792	95
RSIHR	-4.997	1.807	1.791	0.991	0.003	-0.062	6.476	0.794	95
RSIHR2	-5.000	1.807	1.765	0.976	-0.000	0.001	6.386	0.796	96
R.corr	-4.997	1.808	1.791	0.991	0.003	-0.056	6.480	0.794	95

Table 2.7: Contraceptive study: inferential statistics for various designs testing $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = -5$ in 10,000 iterations, $\theta^T = (127, 132, 330, 235)$, $\rho = 0.05$ and $n = 355$.

has lower MSE for both the null and alternative hypothesis, which is not surprising because its variability of probability of assigning to the experimental arm is limited by its block size of 8. Under the null hypothesis, Neyman allocation has the lowest bias, and is able to remain under the nominal 5% alpha level. However, Neyman allocation does not perform as well relative to R.corr under the alternative hypothesis, with the former having relative bias of -0.177 and power of 0.792, and the latter having a smaller relative bias of -0.056 and a higher power of 0.794. However, it is noted that RSIHR2 allocation, which seeks to minimize the total number of failures (or, minimize the total expected response in the continuous case) and to place less subjects in the inferior arm, performs best under the alternative hypothesis, with relative bias of 0.001, highest power among the response-adaptive designs of 0.796, and the highest empirical coverage of 96%. The increased power and empirical coverage can be attributed to RSIHR2's SD/SE ratio being the lowest here at 0.976, pointing to an overestimation of the standard error of the treatment effect estimate. The SD/SE provided by R.corr and RSIHR under the alternative hypothesis are closer to 1 than that of the other RAR designs, indicating low biases and reasonably estimated standard errors of the treatment effect estimate.

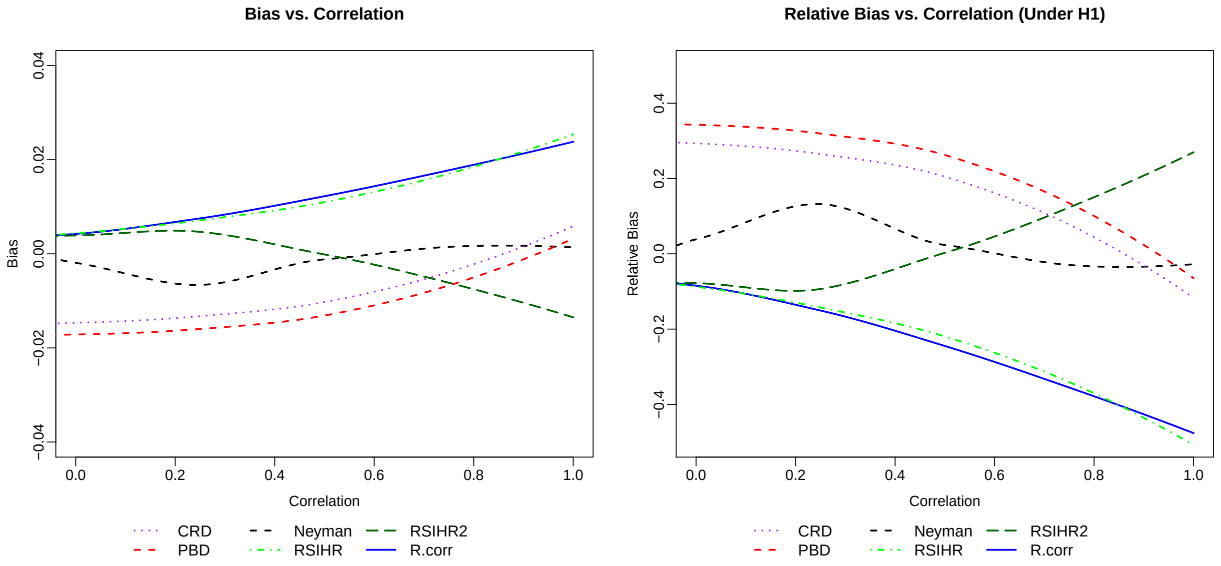
When higher correlation is introduced into the contraceptive study simulations, with $\rho = 0.30$, (Table 2.8), the importance of adjusting for correlation is evident, especially under the alternative hypothesis. Under the null, RSIHR2 has the lowest bias of -0.002, compared to R.corr's bias of -0.007. RSIHR2's low bias under the null and low standard error for the treatment effect estimate results in lowest MSE. With the exception of RSIHR2, R.corr still outperforms the remaining designs in terms of bias (-0.007), and is able to maintain Type I error under the 5% nominal alpha level, in spite of overestimation of the standard error of the treatment effect estimate (lowest sd/se ratio under the null amongst designs considered). On the other hand, CRD's Type I error slips just above 5%, due to its higher bias and standard error.

	$\hat{\beta_1}$	se($\hat{\beta_1}$)	sd	sd_se	bias	rb	mse	size	ci
Under H_0									
CRD	-0.013	1.785	1.777	0.996	-0.013		6.349	0.051	95
PBD	-0.015	1.783	1.774	0.995	-0.015		6.330	0.049	95
Neyman	-0.008	1.805	1.773	0.982	-0.008		6.404	0.047	95
RSIHR	-0.009	1.804	1.770	0.981	-0.009		6.391	0.046	95
RSIHR2	-0.002	1.792	1.765	0.985	-0.002		6.333	0.046	95
R.corr	-0.007	1.809	1.773	0.980	-0.007		6.423	0.045	95
Under H_1									
CRD	-5.013	1.785	1.777	0.996	-0.013	0.255	6.349	0.802	95
PBD	-5.015	1.783	1.774	0.995	-0.015	0.310	6.330	0.802	95
Neyman	-5.015	1.805	1.779	0.986	-0.015	0.304	6.427	0.793	95
RSIHR	-4.991	1.807	1.795	0.993	0.009	-0.180	6.490	0.792	95
RSIHR2	-4.952	1.805	1.766	0.979	0.048	-0.960	6.384	0.787	95
R.corr	-4.992	1.812	1.800	0.993	0.008	-0.166	6.525	0.790	95

Table 2.8: Contraceptive study: inferential statistics for various designs testing $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = -5$ in 10,000 iterations, $\theta^T = (127, 132, 330, 235)$, $\rho = 0.30$ and $n = 355$.

Under the alternative hypothesis, the benefits of adjusting for correlation are more evident, as can be seen by lowest relative bias. We note that R.corr's power of 0.790 is not as strong as Neyman allocation's power of 0.793, yet R.corr's relative bias is much lower at -0.166 compared to Neyman's relative bias of 0.304. It is also worth noting the high relative bias of -0.960 of RSIHR2, showing that at this level of correlation, RSIHR2's estimate of the treatment effect is underestimated to a larger degree than that of the other designs. RSIHR2 has the lowest power of 0.787, explained by large bias and overestimation of the standard error. The SD/SE ratio close to 1 of R.corr under the alternative hypothesis indicates reasonable estimates of standard error, and gives confidence in the strong performance with respect to biases, power, and empirical coverage.

The same simulations were run for other values of ρ. The results of bias and relative bias under the alternative hypothesis versus correlation are shown in Figures 2.9a and 2.9b. It is interesting to note that the curves plotting bias of the treatment effect estimate under Complete Randomized Design, Permuted Block Design, RSIHR, and R.corr share similar shapes. The very close performance of RSIHR allocation and R.corr should not be surprising since their derivations both depend on minimizing total expected responses, albeit RSIHR does not adjust for correlation. While RSIHR and RSIHR2 curves begin closely together, when correlation exceeds 0.25, the RSIHR2 bias begins to decrease and continues to do so as correlation increases. The deviation in behavior is due to RSIHR2's adjustment of ensuring that the total expected response is minimized *and* less patients are placed in the inferior arm. With 5% correlation, as shown in Table 2.7, the relative bias under the alternative hypothesis of RSIHR2 is the lowest at 0.001, while that of R.corr is next with relative bias of -0.056. Neyman allocation has relative bias of -0.177, and PBD performs worst in this metric with relative bias of 0.340. However, as can be seen in Figures 2.9a and 2.9b, Neyman allocation performs increasingly poorly as correlation increases from 0 to 0.25, whereby its performance improves as



(a)

(b)

Figure 2.9: Contraceptive study: bias and relative bias under the alternative hypothesis for varying levels of correlation.

correlation increases from 0.25 to 1.

Power as a function of correlation for the contraceptive study is shown in Figure 2.10. The power of R.corr declines as correlation increases. The loss in power results from the increasing number of patients allocated to treatment group E, leading to larger treatment group imbalance as correlation increases, as well as increased variability in n_E/n (see Table 2.9) as correlation increases. The faster decline of power in R.corr design as correlation increases is due to the larger increase in variability of the target allocation $R = n_E/n_C$. For example, as correlation increases from 0.5 to 0.55, the variability of R in RSIHR design remains consistent around 0.05, but the variability of R in R.corr design increases from 0.351 to 0.523.

It is suggested to increase the sample size as correlation increases and still utilize the R.corr allocation, due to its demonstrated ability to minimize the total response under the alternative, and its ability to more accurately estimate the treatment effect (e.g. see relative bias in Table 2.8). For example, in the contraceptive example with $\rho = 0.05$, a mere increase of sample size from 355 to 363 was sufficient to bring power from 79.4 to 80.3%. Similarly, with $\rho = 0.30$, an increase of sample size from 355 to 368 increased power from 79 to 81 %. Meanwhile, the bias under H_0 and relative bias under H_1 were significantly lower under R.corr than other allocation targets. Table 2.10 below shows the resulting power from upwards sample size adjustments to address the drop in power when using R.corr. Note that 475 subjects are needed for 90% power under CRD, yet 480 subjects are needed for 90% power under R.corr.

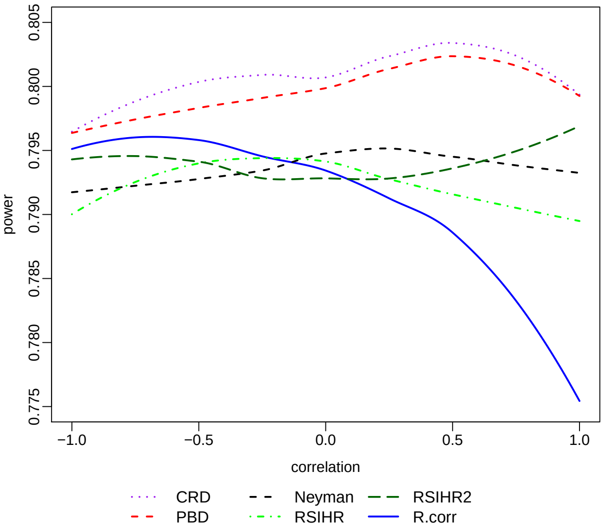


Figure 2.10: Contraceptive study: power vs correlation.

ρ	Neyman	RSIHR	RSIHR2	R.corr
0.00	0.056	0.062	0.038	0.062
0.05	0.056	0.063	0.054	0.071
0.10	0.047	0.064	0.043	0.100
0.15	0.051	0.063	0.034	0.119
0.20	0.056	0.064	0.039	0.130
0.25	0.057	0.074	0.040	0.265
0.30	0.053	0.060	0.041	0.199
0.35	0.051	0.055	0.042	0.185
0.40	0.049	0.054	0.048	0.229
0.45	0.053	0.055	0.037	0.323
0.50	0.057	0.055	0.041	0.351
0.55	0.052	0.053	0.042	0.523
0.60	0.052	0.053	0.039	0.697
0.65	0.051	0.054	0.053	0.753
0.70	0.056	0.058	0.042	1.379
0.75	0.142	0.062	0.045	1.320
0.80	0.057	0.069	0.037	2.116
0.85	0.057	0.061	0.040	2.184
0.90	0.066	0.055	0.037	2.052
0.95	0.052	0.052	0.063	2.959
1.00	0.047	0.055	0.035	3.593

Table 2.9: Contraceptive study: variance of $\hat{R}$ amongst 355 subjects for varying levels of correlation, averaged across 10,000 iterations.

n	CRD	PBD	DBCD.Neyman	RSIHR	RSIHR2	R.corr
365	0.811	0.810	0.807	0.802	0.795	0.798
375	0.823	0.826	0.811	0.811	0.821	0.810
385	0.833	0.837	0.823	0.820	0.825	0.820
395	0.838	0.840	0.833	0.834	0.841	0.833
405	0.850	0.846	0.842	0.839	0.846	0.840
415	0.860	0.863	0.854	0.853	0.851	0.851
425	0.864	0.864	0.866	0.855	0.865	0.852
435	0.867	0.870	0.875	0.868	0.864	0.868
445	0.884	0.879	0.883	0.874	0.882	0.873
455	0.888	0.889	0.883	0.877	0.884	0.877
465	0.890	0.895	0.896	0.893	0.885	0.890
475	0.900	0.895	0.894	0.893	0.898	0.893
480	0.902	0.900	0.901	0.899	0.899	0.900

Table 2.10: Contraceptive study: sample size adjustments and resulting levels of power ranging from 0.8 to 0.9 with $\theta^T = (127, 132, 330, 235)$ and $\rho = 0.30$.

The scleroderma example also shows the value of accounting for correlation when the objective is to minimize the total expected response. It also shows the value of simulation while selecting a study design, as the designs evaluated perform differently than in the contraceptive study. In the study of scleroderma patients, $\mu_E = 21.8$, $\mu_C = 27.5$, $\sigma_E^2 = 219$, $\sigma_C^2 = 144$. The optimal allocation proportion accounting for correlation is shown for varying levels of correlation in Table 2.11. The optimal value of R ranges from 1.3947 when correlation is 0, to 1.5208 when correlation is 1.

ρ	R.corr	ρ	R.corr
0.00	1.3947	0.50	1.4356
0.05	1.3978	0.55	1.4413
0.10	1.4011	0.60	1.4475
0.15	1.4045	0.65	1.4541
0.20	1.4082	0.70	1.4613
0.25	1.4121	0.75	1.4691
0.30	1.4162	0.80	1.4776
0.35	1.4206	0.85	1.4869
0.40	1.4252	0.90	1.4971
0.45	1.4302	0.95	1.5084
		1.00	1.5208

Table 2.11: Scleroderma study: optimal allocation proportion for varying levels of correlation, $\theta^T = (21.8, 27.5, 219, 144)$ and $n = 165$.

After successfully enrolling a total of n patients, we observe the proportion of patients enrolled in the experimental arm under the alternative hypothesis, depicted in Figure 2.11. Similar to Figure 2.7, we can see that the proportion of patients assigned to the experimental arm (group E) for designs that do not adjust for correlation are consistent regardless of correlation between the responses of the two arms. On the other hand, the blue solid line depicts the proportion of patients in group E under R.corr, which places an

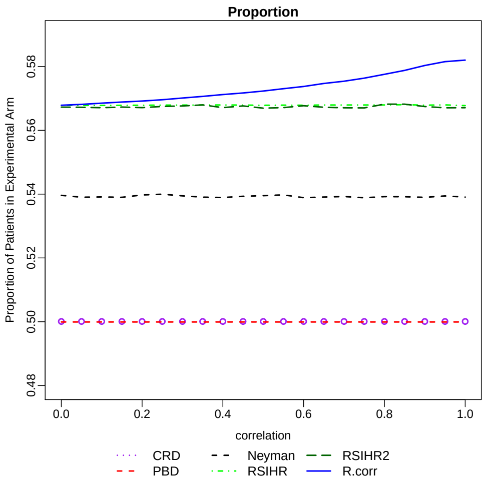
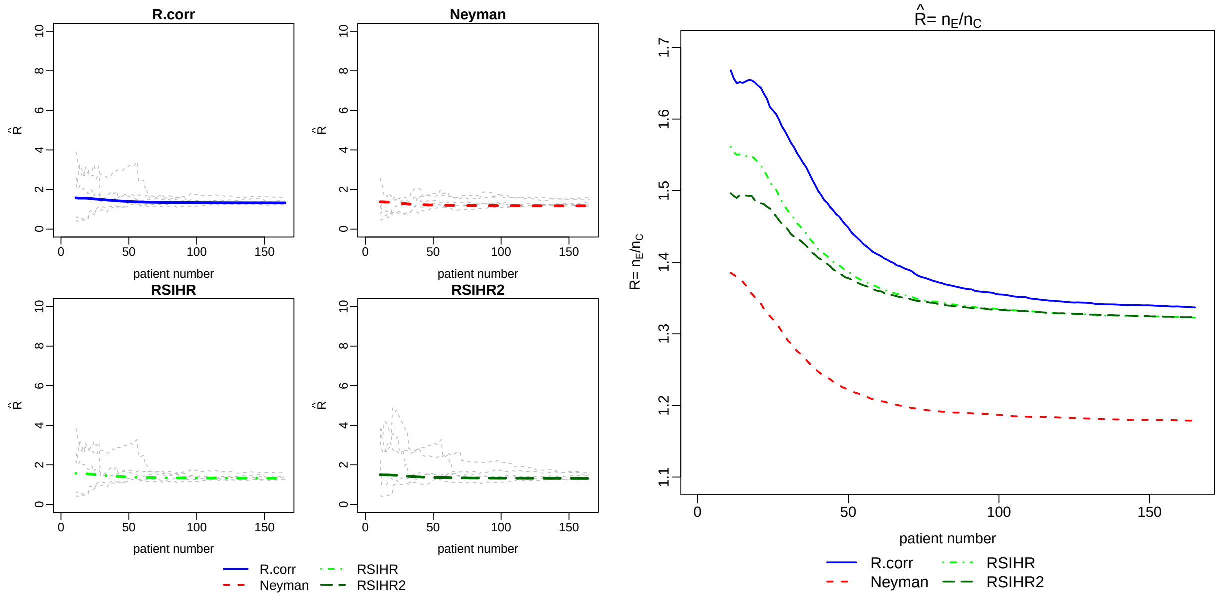


Figure 2.11: Scleroderma study: proportion of patients in the experimental arm (treatment group E) for different levels of correlation, averaged across 10,000 iterations.

increasing proportion of patients in the experimental arm as correlation increases.

The simulation is set up in the same manner as for the contraceptive example. Figure 2.12a shows the value of $\hat{R}$ from patients 1 to $n = 165$ for the doubly-biased coin design (DBCD) targeting R.corr, Neyman, RSIHR, and RSIHR2 allocations. In each quadrant, the dotted gray lines are the value of $\hat{R}$ for a given subject in one simulated trial (a single iteration). The colored lines within the plot show the average $\hat{R}$ across patient enrollment averaged across 10,000 iterations. These average $\hat{R}$ values are shown again in Figure 2.12b for ease of comparison. Note that the larger effect size of the scleroderma example results in the curves for average $\hat{R}$ in RSIHR and RSIHR2 designs to come together more quickly; by patient 60 the two $\hat{R}$'s are in close agreement, compared to this consistency not occurring until after 100 patients were enrolled in the contraceptive example. The larger effect size of the scleroderma case also results in a larger difference in $\hat{R}$ for R.corr versus Neyman designs.

The value of R.corr allocation is evident in Table 2.12, which exhibits total observed response $\bar{Y}_E n_E + \bar{Y}_C n_C$ for $\rho \in (0.05, 0.30, 0.60)$ for the scleroderma case. It can be seen that for all cases, the total response under the *R.corr* design is indeed lower than those of the other designs. Specifically, the comparison of R.corr with RSIHR makes most sense, since both target allocations seek to minimize the total expected response for a fixed variance, with R.corr adjusting for correlation and RSIHR assuming independence. When $\rho = 0.05$ in the scleroderma study, the total response under R.corr is 4022.665, while under RSIHR is 4025.571. The further improvement as gleaned by the size of the gap between the total expected response under R.corr and RSIHR is more apparent as correlation increases; when $\rho = 0.30$, R.corr's total response is 4022.665, compared with RSIHR's 4025.571. Finally, when $\rho = 0.60$, R.corr's total response is 4018.457, compared



(a)

(b)

Figure 2.12: Scleroderma study: (a) $\hat{R} = n_E/n_C$ across 165 enrolled patients, $\rho = 0.05$. Each gray dotted line represents $\hat{R}$ in a single iteration. The bolder lines in color represent the average $\hat{R}$ for the design across 10,000 iterations.
(b) $\hat{R} = n_E/n_C$ across 165 enrolled patients averaged across 10,000 iterations.

with RSIHR's total response of 4025.627.

While total response is successfully reduced when using R.corr compared to CRD, PBD, and Neyman, RSIHR, and RSIHR2 targeted by DBCD, we also evaluate the performance of these designs by looking at bias, relative bias, Type I error, and power. Inferential results for the scleroderma study are shown in Table 2.13. Under the null hypothesis, R.corr maintains bias under the 5% nominal alpha level and has bias levels similar to the other designs evaluated. Under the alternative, the relative bias of R.corr is lower than the others, at -1.781, versus a relative bias of -1.810 in the RSIHR design. The power under the alternative drops to 77.4%.

In the scleroderma example, when $\rho = 0.30$ (Table 2.13), R.corr does not differentiate itself much from RSIHR design. For example, the relative bias of RSIHR and R.corr are -1.800 and -1.789, respectively, and powers are similar at 77.4% and 77.5%, respectively. R.corr still outshines CRD, PBD, and Neyman allocation as targeted by DBCD, since those designs had relative bias all more severe than -2.

In the last scenario assessed for the scleroderma study, a high level of correlation of $\rho = 0.60$ is introduced. Under the null hypothesis, R.corr controls its Type I error under 5%, and has the highest empirical coverage of 95.18%. Under the alternative, we see the high correlation is associated with a larger drop in power to 77.3%, however R.corr boasts the smallest relative bias of -1.599, compared with the closest second provided

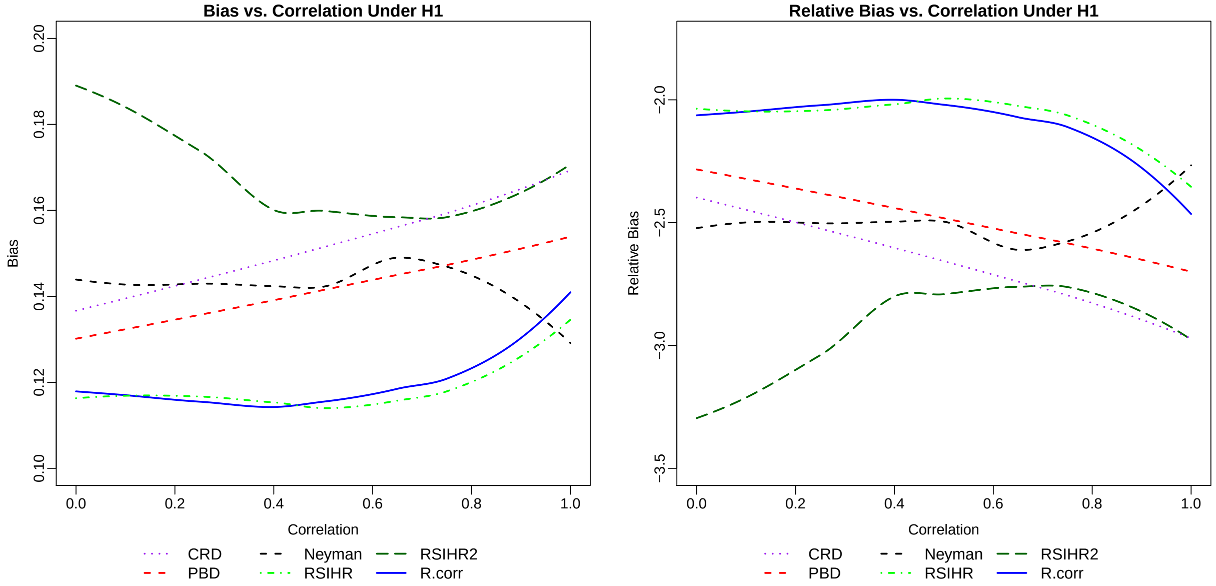
	Patients in E (H1)			**Responses**		**Total Response**
	mean	sd	proportion	μ_E	μ_C	$\mu_E n_E + \mu_C n_C$
$\rho = 0.05$						
CRD	82.516	6.495	0.500	21.986	27.547	4086.427
PBD	82.489	0.962	0.500	21.976	27.545	4085.566
Neyman	88.939	5.685	0.539	21.966	27.544	4048.674
RSIHR	93.686	5.716	0.568	21.980	27.577	4025.877
RSIHR2	93.589	5.611	0.567	22.009	27.560	4027.836
R.corr	93.740	5.808	0.568	21.977	27.575	4025.151
$\rho = 0.30$						
CRD	82.516	6.495	0.500	21.986	27.540	4085.830
PBD	82.489	0.962	0.500	21.976	27.539	4085.108
Neyman	89.009	5.612	0.539	21.973	27.518	4046.900
RSIHR	93.703	5.702	0.568	21.979	27.576	4025.571
RSIHR2	93.659	5.725	0.568	21.983	27.580	4026.523
R.corr	94.069	6.427	0.570	21.974	27.571	4022.665
$\rho = 0.60$						
CRD	82.516	6.495	0.500	21.986	27.531	4085.067
PBD	82.489	0.962	0.500	21.976	27.532	4084.525
Neyman	88.915	5.591	0.539	21.954	27.527	4046.479
RSIHR	93.696	5.714	0.568	21.977	27.578	4025.627
RSIHR2	93.670	5.640	0.568	21.991	27.593	4028.047
R.corr	94.666	7.955	0.574	21.964	27.572	4018.457

Table 2.12: Scleroderma study: total observed response under the alternative hypothesis $H_1 : \beta_1 = -5.7$; $\theta^T = (21.8, 27.5, 219, 144)$ and $n = 165$.

by RSIHR2 at -1.714. While RSIHR2 seems to be a consistently strong perform, remember in the scleroderma study with $\rho = 0.05$ that RSIHR2 had the highest relative bias amongst the designs assessed. Here with a correlation of 0.60, CRD had the highest relative bias under the alternative at -2.712.

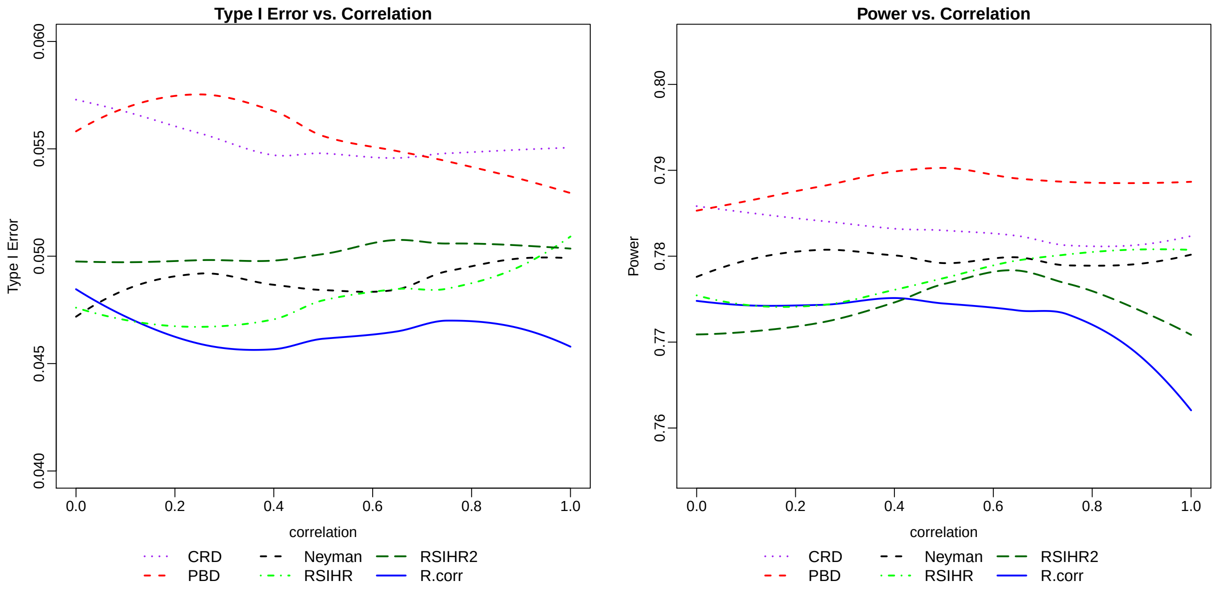
Figures 2.13a and 2.13b plot bias and relative bias versus correlation. Note that the shapes of these curves look very different from those provided by the contraceptive study in Figures 2.9a and 2.9b. This shows the importance of simulation to prepare expectations prior to trial implementation. In both Figures 2.13a and 2.13b, we see that R.corr and RSIHR have the lowest biases and relative biases in most cases of correlation (except when correlation is nearly 1, when Neyman allocation has lower bias and relative bias). It is also interesting to note that PBD has lower bias and relative bias than CRD, regardless of correlation level.

Figures 2.14a and 2.14b plot Type I error and Power versus correlation. In assessing Type I error, we see that R.corr consistently controls Type I error, while CRD and PBD often times are unable to do so. Neyman



(a) (b)

Figure 2.13: Scleroderma study: bias and relative bias under the alternative hypothesis for varying levels of correlation.



(a) (b)

Figure 2.14: Scleroderma study: Type I error and power for varying levels of correlation.

	$\hat{\beta_1}$	se($\hat{\beta_1}$)	sd	sd_se	bias	rb	mse	size	ci
Under H_0									
CRD	0.1381	2.059	2.070	1.005	-0.000		4.241	0.057	94.27
PBD	0.1313	2.052	2.061	1.004	0.000		4.213	0.056	94.37
Neyman	0.1436	2.086	2.057	0.986	0.000		4.350	0.048	95.25
RSIHR	0.1452	2.086	2.051	0.983	0.000		4.351	0.047	95.27
RSIHR2	0.1608	2.070	2.042	0.987	0.000		4.283	0.048	95.22
R.corr	0.1419	2.087	2.055	0.985	0.000		4.355	0.049	95.12
Under H_1									
CRD	-5.5613	2.016	2.027	1.006	0.138	-2.423	4.127	0.785	94.18
PBD	-5.5688	2.009	2.017	1.004	0.131	-2.291	4.086	0.786	93.93
Neyman	-5.5771	2.032	1.986	0.977	0.122	-2.146	3.960	0.783	95.11
RSIHR	-5.5963	2.053	2.025	0.986	0.103	-1.810	4.111	0.775	95.00
RSIHR2	-5.5510	2.054	2.026	0.987	0.148	-2.604	4.129	0.774	94.56
R.corr	-5.5979	2.054	2.026	0.986	0.102	-1.781	4.114	0.774	94.89

Table 2.13: Scleroderma study: inferential statistics for various designs testing $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = -.5.7$ in 10,000 iterations evaluating $\theta^T = (21.8, 27.5, 219, 144)$, $\rho = 0.05$ and $n = 165$.

	$\hat{\beta_1}$	se($\hat{\beta_1}$)	sd	sd_se	bias	rb	mse	size	ci
Under H_0									
CRD	0.1453	2.060	2.069	1.005	0.000		4.242	0.055	94.51
PBD	0.1368	2.052	2.056	1.002	0.000		4.213	0.057	94.29
Neyman	0.1261	2.087	2.063	0.988	0.000		4.357	0.048	95.25
RSIHR	0.1532	2.086	2.053	0.984	0.000		4.350	0.047	95.33
RSIHR2	0.1060	2.070	2.045	0.988	-0.000		4.285	0.048	95.25
R.corr	0.1431	2.095	2.059	0.983	0.000		4.388	0.046	95.37
Under H_1									
CRD	-5.5541	2.016	2.026	1.005	0.145	-2.550	4.124	0.783	94.06
PBD	-5.5633	2.009	2.013	1.002	0.136	-2.389	4.069	0.787	93.98
Neyman	-5.5454	2.031	2.024	0.996	0.154	-2.703	4.119	0.771	94.47
RSIHR	-5.5969	2.054	2.033	0.990	0.103	-1.800	4.142	0.774	94.88
RSIHR2	-5.5971	2.052	2.021	0.985	0.102	-1.795	4.095	0.777	94.74
R.corr	-5.5974	2.057	2.027	0.985	0.102	-1.789	4.118	0.775	94.76

Table 2.14: Scleroderma study: inferential statistics for various designs testing $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = -.5.7$ in 10,000 iterations evaluating $\theta^T = (21.8, 27.5, 219, 144)$, $\rho = 0.30$ and $n = 165$.

allocation is able to control Type I error at the nominal 5% for small to moderate levels of correlation, but after correlation surpasses 0.50, the ability of Neyman allocation to control Type I error is diminished. In Figure 2.14b, R.corr has stronger power than Neyman allocation for small levels of correlation, but after correlation exceeds 0.4, the power of a design targeting R.corr falls more dramatically. The lower power of R.corr relative to other designs when correlation exceeds 0.4 is due to the higher variability of $\hat{R} = n_E/n_C$, and the faster decline in power is a due to the larger increase variability of the design as correlation increases (Table 2.16).

While 165 patients is an adequate sample size for 80% power in the independent response case, we see from Table 2.14 that power had been reduced to 77.1% for DBCD targeting Neyman allocation and 78.7%

	$\hat{\beta_1}$	se($\hat{\beta_1}$)	sd	sd_se	bias	rb	mse	size	ci
Under H_0									
CRD	0.1546	2.059	2.069	1.005	-0.000		4.241	0.055	94.46
PBD	0.1439	2.052	2.051	0.999	0.000		4.212	0.055	94.52
Neyman	0.1464	2.084	2.057	0.987	-0.000		4.341	0.050	94.98
RSIHR	0.1469	2.086	2.061	0.988	0.000		4.350	0.049	95.09
RSIHR2	0.1432	2.068	2.078	1.004	-0.000		4.278	0.050	94.95
R.corr	0.1536	2.114	2.079	0.983	-0.000		4.471	0.048	95.18
Under H_1									
CRD	-5.5448	2.016	2.025	1.005	0.155	-2.712	4.125	0.783	94.18
PBD	-5.5562	2.009	2.008	0.999	0.143	-2.513	4.051	0.789	94.25
Neyman	-5.5733	2.032	1.989	0.979	0.126	-2.212	3.972	0.779	95.01
RSIHR	-5.6011	2.053	2.022	0.985	0.098	-1.725	4.099	0.779	94.53
RSIHR2	-5.6018	2.053	2.016	0.982	0.098	-1.714	4.074	0.779	94.89
R.corr	-5.6083	2.066	2.041	0.988	0.091	-1.599	4.173	0.773	94.78

Table 2.15: Scleroderma study: inferential statistics for various designs testing $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = -.5.7$ in 10,000 iterations evaluating $\theta^T = (21.8, 27.5, 219, 144)$, $\rho = 0.60$ and $n = 165$.

ρ	Neyman	RSIHR	RSIHR2	R.corr
0.00	0.100	0.114	0.091	0.114
0.05	0.080	0.112	0.092	0.126
0.10	0.084	0.122	0.104	0.138
0.15	0.085	0.119	0.094	0.162
0.20	0.078	0.113	0.090	0.186
0.25	0.085	0.110	0.090	0.212
0.30	0.083	0.107	0.094	0.235
0.35	0.090	0.104	0.093	0.278
0.40	0.076	0.102	0.104	0.311
0.45	0.073	0.108	0.104	0.365
0.50	0.073	0.101	0.090	0.398
0.55	0.087	0.103	0.095	0.463
0.60	0.080	0.105	0.089	0.609
0.65	0.081	0.113	0.091	0.789
0.70	0.092	0.127	0.094	0.935
0.75	0.079	0.117	0.089	1.052
0.80	0.075	0.116	0.094	1.364
0.85	0.099	0.119	0.095	1.595
0.90	0.077	0.111	0.086	1.607
0.95	0.079	0.105	0.087	1.771
1.00	0.083	0.095	0.090	2.710

Table 2.16: Scleroderma study: variance of $\hat{R}$ amongst 165 subjects for varying levels of correlation, averaged across 10,000 iterations.

for PBD. We also observed in Figure 2.14b that when using R.corr, power drops more drastically than the other designs considered as correlation increases. An upward adjustment needs to be made on sample size in the presence of correlated responses between treatment arms. Table 2.17 shows power with increasing sample sizes from simulations with responses between treatment arms having a correlation of 30%. An increase in sample size to 185 results in 82.09% power in R.corr. While 212 patients would have been a sufficient sample

n	CRD	PBD	Neyman	RSIHR	RSIHR2	R.corr
165	0.7835	0.7873	0.7711	0.7742	0.7772	0.7749
185	0.8226	0.8258	0.8226	0.8197	0.8185	0.8209
205	0.8591	0.8612	0.8637	0.8582	0.8503	0.8550
225	0.8909	0.8912	0.8914	0.8946	0.8900	0.8918
245	0.9171	0.9155	0.9167	0.9100	0.9149	0.9101

Table 2.17: Scleroderma study: sample size adjustments and resulting levels of power ranging from 0.8 to 0.9 ($\theta^T = (21.8, 27.5, 219, 144)$ and $\rho = 0.30$).

size for 90% power given independent responses, we can see from the Table that a sample size greater than 225 is needed when correlation is 30%.

2.4.2 Nonparametric Methods in Assessing Performance of Optimal Allocation Proportion Accounting For Correlation

Proschan stated that analysis of trial data can ignore correlation between responses if the randomization scheme also did not account for the correlation [69]. Tables 2.7- 2.8, and 2.13 - 2.15 show five randomization schemes that ignore correlation, and one randomization scheme that targets an optimal allocation proportion that accounts for the correlation. One might argue that the comparison is unfair. To address this issue, we can use nonparametric methods to compare the six randomization schemes on a level playing field.

Bootstrap is a nonparametric simulation technique that estimates the sampling distribution of a statistic by simulation. The method draws a sample of original size n_E and n_C from the data by sampling with replacement. By doing so, the sample is treated as the whole population, and the resampling allows one to study the sampling distribution of any statistic, such as the sample mean, or in our case, the sample difference in means between two treatment groups. The theory behind the bootstrap relies on asymptotics: as the sample size of the original data goes to infinity, the more accurate the inferences drawn from this method [28]. In general, 500 to 1000 bootstraps is often recommended.

In each iteration of the simulation in section 2.4.1, y_E and y_C are treated as responses from the entire population. There are n_E observations drawn from the E population with replacement, and n_C observations drawn from the C population with replacement. Each such occurrence is considered a single bootstrap which yields a single statistic, and 1000 bootstraps are performed. The statistic we are interested in is $(\bar{Y}_E - \bar{Y}_C)$. With each iteration of the simulation, 1000 bootstraps yields 1000 sample statistics. We then compute the 95% confidence interval from the 2.5^{th} and 97.5^{th} quantiles of the 1000 sample statistics. Since the simulation includes 10000 iterations for each randomization design, we have 10000 confidence intervals for each randomization design as well. The widths of these intervals shed light on our certainty of the estimate of the difference in the means in the two groups. The confidence intervals from the bootstrap method is displayed in Tables 2.18 for the contraceptive trial and 2.19 for the scleroderma trial.

	$\rho = 0.05$			$\rho = 0.30$		
	$\hat{\beta}_1$	95% CI	Width	$\hat{\beta}_1$	95% CI	Width
CRD	-5.000	[-9.8097, -0.1885]	9.6212	-5.001	[-9.5052, -0.4958]	9.0094
PBD	-4.999	[-9.8505, -0.1543]	9.6962	-5.001	[-9.5353, -0.4658]	9.0695
Neyman	-5.000	[-9.7569, -0.2438]	9.5131	-5.000	[-9.4552, -0.5415]	8.9136
RSIHR	-5.000	[-9.7464, -0.2523]	9.4941	-5.000	[-9.4502, -0.5527]	8.8975
RSIHR2	-5.000	[-9.7449, -0.2564]	9.4885	- 5.001	[-9.4492, -0.5491]	8.9001
R.corr	-5.001	[-9.7475, -0.2546]	9.4929	-5.000	[-9.4395, -0.5583]	8.8812

Table 2.18: Contraceptive study: average treatment effect estimate $\hat{\beta}_1$ and 95% confidence intervals from 1000 bootstraps for each of 10,000 iterations.

Table 2.18 shows the confidence intervals and widths from the 1000 bootstraps for the contraceptive study. It can be seen that when $\rho = 0.05$ and $\rho = 0.30$, the confidence intervals of the difference in means in the design targeting R.corr are more narrow than those of the designs targeting Neyman and RSIHR allocations. The narrowest width when $\rho = 0.05$ is from RSIHR2, which coincides with its largest empirical coverage of 96% in the parametric analyses as shown in Table 2.7. When $\rho = 0.30$, R.corr yields the most narrow confidence interval of all six designs, demonstrating higher certainty of estimates of R.corr. The confidence interval narrows from Neyman to R.corr allocations for $\rho = 0.05$ and $\rho = 0.30$, with the widths for Neyman allocation being 9.5131 and 8.9136, respectively, and the widths for R.corr being 9.4929 and 8.8812 respectively. We observe that the confidence interval width decreases as correlation increases.

	$\rho = 0.05$			$\rho = 0.30$			$\rho = 0.60$		
	$\hat{\beta}_1$	95% CI	Width	$\hat{\beta}_1$	95% CI	Width	$\hat{\beta}_1$	95% CI	Width
CRD	-5.5598	[-11.1872, 0.0698]	11.2570	-5.5545	[-10.8317, -0.2809]	10.5508	-5.5437	[-10.3594, -0.7236]	9.6358
PBD	-5.5705	[-11.2563, 0.1192]	11.3755	-5.5627	[-10.8802, -0.2454]	10.6349	-5.5567	[-10.3988, -0.7155]	9.6833
Neyman	-5.5780	[-11.1619, 0.0073]	11.1691	-5.5457	[-10.7801, -0.3126]	10.4675	-5.5715	[-10.3574, -0.7877]	9.5697
RSIHR	-5.5942	[-11.1222, -0.0679]	11.0543	-5.5965	[-10.7865, -0.4038]	10.3827	-5.6011	[-10.3571, -0.8465]	9.5106
RSIHR2	-5.5487	[-11.0736, -0.021]	11.0526	-5.5972	[-10.7910, -0.4038]	10.3872	-5.5998	[-10.3547, -0.8463]	9.5085
R.corr	-5.5986	[-11.1268, -0.0727]	11.0541	-5.5977	[-10.7885, -0.4071]	10.3814	-5.6074	[-10.3677, -0.8496]	9.5180

Table 2.19: Scleroderma study: average treatment effect estimate $\hat{\beta}_1$ and 95% Confidence Intervals from 1000 bootstraps for each of 10,000 iterations.

In Table 2.19, the confidence intervals and widths resulting from 1000 bootstraps is shown for the scleroderma study example. The confidence interval narrows from Neyman to R.corr allocations for $\rho = 0.05, 0.30,$ and 0.60, with the widths for Neyman allocation being 11.1691, 10.4675, and 9.5697, respectively, versus 11.0541, 10.3814, and 9.5180 for R.corr, respectively. The estimate $\hat{\beta}_1$ is closer to the true value of $\beta_1 = -5.7$ when using R.corr than the other designs. The widths of the bootstrapped confidence intervals are consistently more narrow for R.corr than those of CRD and PBD. This is consistent with the larger empirical coverage of R.corr as reported in the "ci" column in Tables 2.13 through 2.15. Similar to its performance in the contraceptive study, R.corr has a wider confidence interval than RSIHR when the correlation is 0.05, but a more narrow confidence interval than RSIHR when the correlation is 0.30. The larger width of R.corr's bootstrap confidence interval when correlation is 0.60 relative to those of RSIHR and RSIHR2 coincides with the parametric results of the scleroderma study shown in Figure 2.14b, which depicts RSIHR and RSIHR2 having higher power than R.corr when correlation surpasses 0.4.

2.5 Discussion

In this chapter, *optimal* allocation proportions for clinical trials with correlated continuous responses between two treatment groups were derived, where smaller responses are considered better, and optimal is defined as minimization of the expected total response from two treatment groups for a fixed power. The optimal allocation was shown to be a function of the ratio of the two groups' means and variances, and the degree of correlation between outcomes from the two-arm trial. When instead a large response is desirable, we cannot obtain an optimal allocation by maximizing $\mu_E n_E + \mu_C n_C$ subject to $Var(\hat{\psi}) = V\hat{a}r(\bar{Y}_E - \bar{Y}_C) = K$. Instead, we can minimize $\frac{n_E}{\mu_E} + \frac{n_C}{\mu + C}$ subject to K [50, 96]. While the technical result is interesting, it will benefit from further work that focuses on more granular details of the framework for practical application. These details include accurate estimation of the correlation between arms as well as incorporating within-arm correlation.

This chapter addresses correlation issues during individual randomization to treatment arms. Another approach is to implement clustered randomized trials (CRTs). We use this section to highlight both advantages and drawbacks of CRTs, and how the findings of this Chapter address correlation differently.

CRTs are popular due to their ability to reduce contamination resulting from intermingling of subjects in different treatment arms. Contamination is a concern in trials where patients have some control with respect to their treatment. For example, in a trial investigating the efficacy of mindfulness exercises on hypertension, individuals in the control arm could learn about such exercises and try to adopt them themselves. Cluster randomized trials minimize treatment contamination by allocating an intervention to an entire group of units. For example, a set of communities could be assigned an intervention, while another set could be assigned a control treatment. Another example is randomizing entire family units to a dietary intervention or control group, so that family members cannot discuss varying treatments amongst themselves.

Besides its strength in decreasing risk for contamination, clustered randomized trials are able to address common exposures that lead to violation of the independence assumption, since random effects models can include a random effect for each source of common exposure (e.g. a random effect for hospitals), and observations can be modeled with a covariance structure [43].

In a single random effect model, the correlation between two observations in the same cluster is called the intra-cluster correlation (ICC), which is $\rho_{ICC} = \frac{\tau^2}{\tau^2 + \sigma_w^2}$, where τ^2 is the variance of the random effect for cluster j, and σ_w^2 is the variance of the error term. Thus, in this model, observations within a cluster share a common correlation ρ_{ICC}, which is a shared correlation for both treatment and control clusters.

In the two random effects model, the separate random cluster effects for treatment and control clusters (τ_T^2 and τ_C^2, respectively) results in different ICCs for each treatment group. If we allow subjects within clusters of the control group to share a correlation of ρ_{ICC_C}, and subjects within clusters of the treatment group to share a correlation of ρ_{ICC_T}, then the $\rho_{ICC\_C} = \frac{\tau_C^2}{\tau_C^2+\sigma_w^2}$ and the $\rho_{ICC\_T} = \frac{\tau_T^2}{\tau_T^2+\sigma_w^2}$. Random variation may exist in the control group's clusters (τ_C^2) and in the treatment group's clusters (τ_T^2).

The intracorrelation cluster coefficient (ICC) helps researchers calculate sample size accounting for correlation between clusters, by adjusting the sample size resulting from standard sample size calculations via a multiplicative factor, known as the *variation inflation factor* or *design effect*:

$$1 + (n_{avg} - 1)\rho_{ICC},$$

where n_{avg} is the average cluster size, and ρ_{ICC} is the ICC.

Torgerson questioned whether clustered randomization was the best solution to contamination within trials, and summarized the disadvantages of utilizing the clustered design. First, cluster trials usually require a larger sample size than would be required in similar, individually randomized trials. Second, cluster trials tend to experience recruitment bias, where certain clusters or patients are more likely to be allocated to a certain treatment. Third, although covariate balance at baseline can be achieved if randomization is performed at the cluster level with a sufficient number of clusters, simple randomization of clusters can still result in covariate imbalance. For example, in a randomized trial of breast cancer screening, the study arms witnessed an imbalance in socioeconomic groups, in spite of the inclusion of 87 clusters and 50,000 women [2]. With randomization at the individual level, a sample size this large would have a very low probability of witnessing covariate imbalance. For more on covariate imbalance, refer to Section 4.1.3. Fourth, selection bias is an indismissable risk. Clusters are typically randomized to a treatment, whereby participants are asked post-randomization whether they consent to treatment and inclusion in study analysis (often requiring follow-up visits). If a large proportion of participants do not provide their consent, or if participants are more likely to consent to one treatment type than other, selection bias is a risk (see more on selection bias in Section 4.1.4). Fifth, clustered designs sometimes employ Zelen's method, of which there are two types: single and double consent. In single consent design, patients who rejected the experimental treatment upon initial offering are then offered the control treatment. In double consent design, patients are initially offered the treatment to which they were randomized; if they decline the randomized treatment, they are then offered alternative therapies, including the investigative treatment [83, 84]. Utilizing Zelen's method consequently dilutes the treatment effect and thus requires a further upwards adjustment in sample size.

In addition to Torgerson's examples on why individual randomization rather than clustered randomization ought to be considered, we note that in clustered randomized trials, the random effect that contributes to correlation between responses needs to be identified – is it the treatment site?, the community?, the personnel administering treatment? How can we still address correlation when the source is unknown? While Biswas was the first to utilize individualized randomization while accounting for correlation for clinical trials with binary responses, this chapter expands the work to trials with continuous responses, resulting in the target allocation R.corr. This allocation is considered an extension of RSIHR allocation discussed in Section B, as RSIHR allocation minimized the expected total response in trials with two arms for independent responses, and R.corr minimizes the expected total response while adjusting for correlation between treatment arms. A natural extension of this work would be to incorporate within arm correlation into the minimization of $\mu_E n_E + \mu_C n_C$ while holding the variance of the treatment effect estimate fixed.

Optimal in this work has thus far referred to the minimization of the total expected response for a given power. We have seen that even in the case of $\mu_E = 127$, $\mu_C = 132$, $\sigma_E^2 = 330$, $\sigma_C^2 = 235$, and $\rho = 0.05$, the *optimal* allocation proportion R.corr places $(\frac{0.548}{1-0.548}) = 21\%$ more subjects in the experimental arm. The difference in patient allocation when using R.corr versus other target allocations is more apparent for larger values of correlation: when $\rho = 0.30$, R.corr places $(\frac{0.554}{1-0.554}) = 24\%$ more subjects in the experimental arm. The ability of R.corr to result in lower total response $\mu_E n_E + \mu_C n_C$ is also more apparent for larger values of correlation, as could be seen in Tables 2.6 and 2.12 for contraceptive and scleroderma examples, respectively.

The value of adjusting for correlation is also evident in the lower relative bias of the treatment effect estimate, as was shown in contraceptive and scleroderma examples of Section 2.4.1. Finally, the presence of correlation requires an upwards adjustment of sample size in order to deliver sufficient power, as shown in Tables 2.10 and 2.17.

While there is value in targeting the R that solves Equation 2.2, this optimal allocation proportion may be difficult to implement in practice. An estimate of correlation from prior data requires the data to be paired, and is sensitive to the pairing decisions. An unequal number of subjects in each treatment arm would result in truncation of observations from the more heavily represented arm. In practice, pairing of subjects in different treatment arms is difficult. The subjects could be paired based off of their genetic similarities, or matched on baseline covariates such as time surpassed since screening and infectious disease outbreak, or white blood cell count range and tumor size. Follmann stated one option was to pair subjects by their ranked pairwise correlations [30] calculated from covariance matrices of their baseline covariates.

Another disadvantage of R.corr is the resulting imbalance between the treatment group sizes may lead to other concerns. For example, placing 21% more subjects in experimental arm E leads to a higher likelihood of having a large string of patients consecutively allocated to E, which could negatively impact the power of the trial to detect a treatment effect, and places the design at risk for biases due to time trends. Treatment group size imbalance will be the topic of Section 4.1.1. Furthermore, one may wonder how the imbalanced allocation performs with respect to other common statistical issues such as covariate imbalance and accidental bias, which are the topics of Sections 4.1.3 and 4.1.2, respectively. Indeed, the accidental bias is minimized when treatment group sizes are equal [28]. Thus, while optimal allocation proportions derived in this chapter is able to minimize total expected responses, the ability to hedge against other potential biases is important. This is the motivation for Chapter 4, which discusses how to assess the strengths and weaknesses of various randomization procedures and designs, and presents a framework for clinical trial design selection. In order to discuss this framework and investigate design choices, it is helpful to first review desirability functions, the topic of the following chapter.

Chapter 3

Desirability Functions:

A Literature Review

Chapter 1 provided a brief overview of a select number of clinical trial designs. Chapter 2 expanded on the RSIHR allocation scheme in Section B to find an optimal target allocation for two-arm trials where responses between the arms are correlated. Chapter 2 concluded by showing that R.corr's performance did well with respect to minimizing expected total response, but had varying satisfactory levels of performance depending on the trial data with regards to other design characteristics of interest such as bias and power. With no straight-forward decisions on design selection due to conflicting strengths and weaknesses, the choice of a design to implement can be difficult. To further investigate design choices as we will do later in Chapter 4, it is helpful first to review the main tool we use to accomplish our task: desirability functions.

A desirability score is a continuous measure that takes on values between zero and one, with large values representing greater desirability. Harrington introduced the concept of a desirability function in 1965, and used it to measure the goodness of a product. When the quality of a final product or outcome is determined by several components or subscores, as is often the case, a desirability function is constructed for each component first before they are combined into an overall desirability function. The approach is attractive because it provides a simple way of arriving at an overall assessment of the product based on several - and frequently disparate - measures of quality of the product. One can view desirability functions as a way of handling multiple response variables [40].

Our aims in this chapter are to review the use of desirability functions in the biomedical arena, their recent innovations, and potential use for applications in more challenging problems. Desirability functions' ability to easily capture several components of a product or process in a single score has attracted many users across diverse disciplines since inception, and continues to do so. A Google Scholar search using the keywords "desirability function" yields over 350,000 results, and adding the term "engineering" yields over 100,000, showing its wide use and common acceptance. Here are some of them:

> - industrial engineering: assessing quality of different tire tread compounds (Derringer et al., 1980) [26]; assessing quality of acetic fermentation when using different kinetic parameters (Pizarro, 2003) [67];
>
> -chromatography: evaluating performance of various high-performance liquid chromatograms in the separation of mixtures (Bourguignon, 1991) [16];
>
> - mechanical engineering: rating the ability of varying factor levels to remove material from the surface of a less resistant body using liquid nitrogen as a coolant (Aggarwal, 2008) [1];
>
> - identifying mechanical properties that are associated with higher overall steel quality (Kim et al., 2000) [53];
>
> - agriculture: evaluating quality of callus induction and rating its growth (Honari et al., 2014) [45];
>
> - environmental science: identifying factors that result in the most efficient bioremediation of weathered crude oil in coastal sediment (Mohajeri et al., 2009) [62].

Further evidence of the continuing popularity of desirability functions can be seen in its existence in R packages and popular commercial software Design Expert by StatEase, and JMP by SAS. Given the data, the packages construct individual desirability functions and then the overall desirability function using user-specified options. The software also can find the individual response values that maximize the overall desirability.

Over the years, various improvements on and extensions of desirability functions have been made, and they remain an active topic of research. Such recent improvements include allowing responses to be (i) correlated, (ii) discrete, (iii) collected over time, and (iv) quantified according to the uncertainty associated with the responses used in the desirability function. For example, Wu (2004) defined a new individual desirability function that incorporates variation and correlation among the responses [93]; Coffey et al. (2007) incorporated discrete outcomes [21]; Chen et al. (2015) introduced a desirability function that accounts for repeated measures over time [18], and Chen et al. (2012) incorporated the standard deviation of predicted responses into an augmented desirability function [17].

Interestingly, there is still little use of desirability functions in the biomedical field even though their potential applications seem to be highly relevant, since patients' progress is seldom based on one single measure but based on several submeasures or components. Frequently, problems arise when components that contribute to the score do not provide the same indications or when they are measured on a different scale. For instance, in rheumatoid arthritis, the ACR20 improvement criteria is a binary variable indicating whether there is at least 20% improvement in at least three of the following components: patient assessment, physician assessment, pain scale, disability questionnaire, and acute phase reactant. All five components witnessing a 50% improvement versus three components witnessing a 20% improvement and two components witnessing a decline would by the traditional definition of ACR20 be rated equivalently. Desirability scores can provide an alternative option to capture and evaluate the contributions from various components and help physicians compare their patients' outcomes on a more granular level. There is thus good potential for greater use of desirability functions in biomedical research. Our literature review shows there is increasing use of such functions in recent years, suggesting that the medical research community may be beginning to explore their potential more seriously. Interestingly, our search reveals there are fewer than 20 such papers in biomedical journals with good statistical content. We briefly review these papers and note most of them do not take advantage of the many innovations that have been recently developed for desirability functions.

Section 3.1 reviews the historical development of desirability functions and some of their applications. In Section 3.2, we discuss the existing biomedical applications of desirability functions and its potential for future use. Section 3.3 presents more recent and sophisticated development of desirability functions to tackle increasingly more complex problems. Section 3.4 provides a brief case study showing how desirability functions can help us gain insight in the longitudinal assessment of overall HIV patient status. Section 3.5 introduces algorithms that aid in optimizing the inputs of an overall desirability score.

3.1 Construction of Desirability Functions

3.1.1 The Original Desirability Function (Harrington, 1965)

There are many components that contribute to product quality and these components can vary in scale and relative importance. In particular, some components may present conflicting signals on product quality, complicating its evaluation. Harrington assumed all measures are continuous and proposed to first construct a desirability function using an exponential transformation for each of the components, resulting on scores on a common scale between zero and one, with one being most desirable. An overall desirability score is then constructed by combining the various individual desirability functions [40].

The initial step is to identify which components or responses are of interest to assess the overall quality of a product. After these components are selected, there are two steps to calculate an overall desirability score:

Step 1. Obtain individual desirability scores.

Categorize the bounds of the individual response's acceptable range as one-sided (later designated smaller-the-better or larger-the-better by Derringer et al., 1980) or two-sided (later designated nominal-the-better). One-sided variables work in one direction: values are more desirable either when they are minimized, or when they are maximized. On the other hand, two-sided variables have a target value that is most desirable, and values are less desirable as they deviate away from that target value.

Let y_i be the value of a response. For two-sided variables, let U and L be the upper- and lower- bounds of the outcome's acceptable values, respectively. Let d_0 denote an assigned desirability score to a specific response value y'_{i0} of a two-sided variable, while d_1 and d_2 are assigned desirabilities assigned to two specific response values of a one-sided variable.

For two-sided variables:

$$y'_i = \frac{2y_i - (U + L)}{U - L}.$$

This is followed by the mapping of a single value of y_i0' to an assumed desirability d_0. The desirability of that value should be independent from the values of other responses of interest. For example, one could assign a desirability d_0 of 0.75 to a response y'_{i0}= 0.1, which would mean that a value slightly higher than the midpoint between the upper and lower bounds corresponded to a good desirability score. Then, d_0 and y'_{i0} are used to define a scaling parameter:

$$k = \frac{ln[ln(\frac{1}{d_0})]}{ln|y'_{i0}|}.$$

Then the individual desirability of a two-sided single response is:

$$d_i = exp[-|y'_i|^k].$$

Figure 3.1 shows three desirability functions for percentage weight change with varying scaling parameters.

For one-sided variables: Assign two response values y_{i1} and y_{i2} ($y_{i1} > y_{i2}$) to two desirabilities d_1 and d_2, so that these response values can be rescaled:

$$y'_{ij} = -ln[-ln(d_j)]; j = 1, 2.$$

A linear transformation is then done to derive y'_i:

$$y_i' = \frac{y_i - y_{i2}}{y_{i1} - y_{i2}}(y_{i1}' - y_{i2}') + y_{i2}.$$

Then the individual desirability of the single-bounded variable is:

$$d_i = exp[-exp(-y_i')].$$

Figure 3.2 shows three individual desirability functions for percentage skin score change, where smaller skin scores are better.

Step 2. Obtain an overall desirability score.

After obtaining individual desirability functions for each attribute, an overall desirability score that accounts for the m attributes can be obtained, defined as

$$D = (d_1^{w_1} d_2^{w_2} ... d_m^{w_m})^{\frac{1}{\Sigma_{i=1}^m w_i}}, \tag{3.1}$$

where w_i is the user-specified weight assigned to each individual desirability d_i, with larger weights indicating more importance.

There are several ways to construct an overall desirability function which should have good properties. For example, the overall desirability should increase in value whenever one of the individual desirabilities increases in value. The overall desirability function should also be able to account for the varying amount of contribution in importance to the overall product quality from each component. Harrington argued that individual weights of one were adequate for many scenarios, simplifying the equation to the geometric mean of multiple components' individual desirabilities. Derringer later noted advantages for using weights not necessarily equal to one. Harrington's choice of geometric mean was suitable for industrial manufacturing since low individual desirability scores decrease overall desirability more rapidly than if other functions such as the arithmetic mean were to be used. The geometric mean and the overall desirability score is thus intended to significantly penalize an overall product if even one outcome has an unacceptable or undesired response.

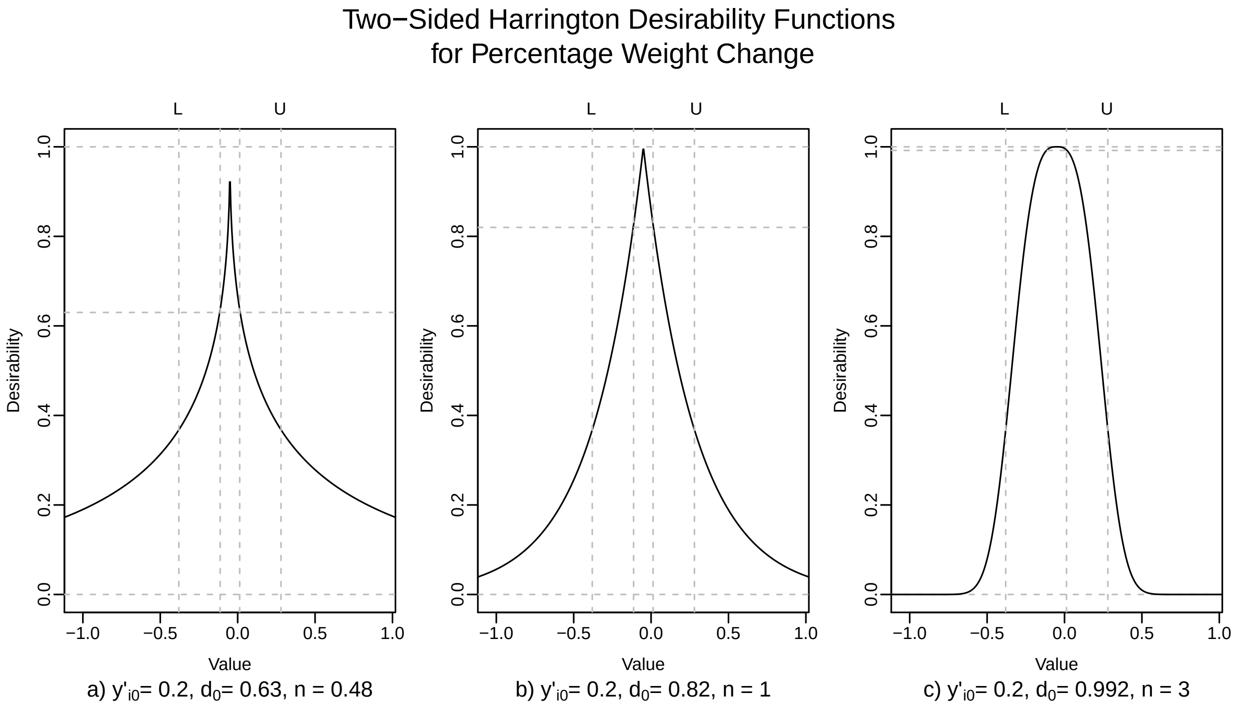
Table 3.1 displays an overall rating system of overall desirability, proposed by Harrington (1965) [40]:

In the literature, the scaling factor k is commonly referred to as a "kurtosis" parameter that a user could specify for two-sided variables, where $k > 1$ results in smoother shapes around the midpoint between L and U, while $k < 1$ penalizes small deviations from this midpoint. The sensitivity to user-specified values d_0

Desirability Score	Interpretation
1.00	The ultimate in satisfaction and quality, where an improvement beyond this point would have no additional meaningful value
[0.80, 1.00)	Acceptable and excellent (represents unusual quality or performance well beyond anything commercially available)
[0.63, 0.80)	Acceptable and good (represents an improvement over the best commercial quality)
[0.40, 0.63)	Acceptable but poor (quality is acceptable to the specification limits but improvement is desired)
[0.30, 0.40)	Borderline (if specification exists, then some of the product quality lies exactly on the specification maximum or minimum)
(0.00, 0.30)	Unacceptable (materials of this quality would lead to failure)
0.00	Completely unacceptable

Table 3.1: Harrington's interpretations of desirability scores.

and y'_{i0} and hence this scaling parameter in two-sided variables can be seen in Figure 3.1, which depicts the desirability function of percentage weight change, with lower specification limit L = -0.38, upper specification limit U = 0.28, and $y'_{i0} = 0.2$.



a) y'_{i0}= 0.2, d_0= 0.63, n = 0.48

b) y'_{i0}= 0.2, d_0= 0.82, n = 1

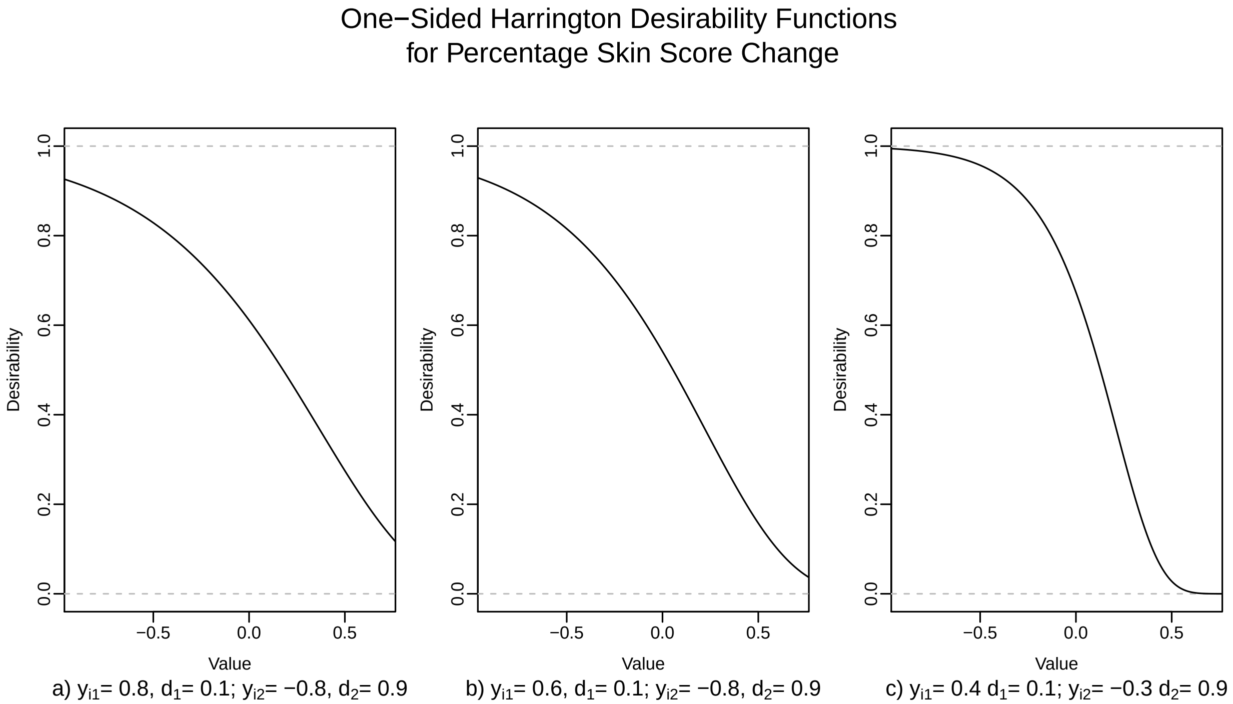
c) y'_{i0}= 0.2, d_0= 0.992, n = 3

Figure 3.1: Harrington two-sided individual desirability functions, with L = -0.38, U = 0.28.

Specifically, in Figure 3.1 a, b, and c, $d_0 = 0.63$, $d_0 = 0.82$, and $d_0 = 0.992$, respectively, resulting in k = 0.48, k = 1, and k = 3, respectively. The interpretation of these values is that a weight percentage change of

-11.3% and +1.3% are a “good” change in Figure 3.1a, an “excellent” change in Figure 3.1b, and a change nearly “ultimate in satisfaction” in Figure 3.1c.

Figure 3.2 depicts the one-sided Harrington desirability functions for percentage skin score change for different values of y_{i1} and y_{i2}. Since smaller values of skin score are considered better, negative percentage skin score change of larger magnitude is desirable. The penalizations as percent skin score change increases progress in levels of severity in Figures 3.2a through c.



a) y_{i1}= 0.8, d_1= 0.1; y_{i2}= −0.8, d_2= 0.9

b) y_{i1}= 0.6, d_1= 0.1; y_{i2}= −0.8, d_2= 0.9

c) y_{i1}= 0.4 d_1= 0.1; y_{i2}= −0.3 d_2= 0.9

Figure 3.2: Harrington one-sided individual desirability functions.

3.1.2 Modified Desirability Function (Derringer et al., 1980)

One would probably agree that the user-specified parameters in Harrington’s method may be a bit confusing and highly subjective. Derringer et al. (1980) modified how the desirability functions are constructed in Step 1 by simplifying the scaling of the variables and making the desirability functions more flexible [26]. Derringer et al.’s simplified method of defining desirability functions is now the default method for constructing desirability functions and is shown below:

Modified Step 1. Obtain Derringer & Suich’s individual desirabilities.

Let y_i, L, and U be defined as before, and T be the target value of a nominal-the-better type response. There are shape parameters r, r_1, and $r_2 > 0$ and they are selected by the user.

For Larger the Better (LTB):

$$d_i = \begin{cases} 0 & \text{for } y_i \leq L \\ (\frac{y_i - L}{U - L})^r & \text{for } L < y_i < U \\ 1 & \text{for } y_i \geq U \end{cases} \tag{3.2}$$

For Smaller the Better (STB):

$$d_i = \begin{cases} 1 & \text{for } y_i \leq L \\ (\frac{U - y_i}{U - L})^r & \text{for } L < y_i < U \\ 0 & \text{for } y_i \geq U \end{cases} \tag{3.3}$$

For Nominal The Better (NTB):

$$d_i = \begin{cases} (\frac{y_i - L}{T - L})^{r_1} & \text{for } L \leq y_i \leq T \\ (\frac{U - y_i}{U - T})^{r_2} & \text{for } T < y_i \leq U \\ 0 & \text{for } y_i < L \text{ or } y_i > U \end{cases} \tag{3.4}$$

We observe that unacceptable values of a response y_i yield an individual desirability d=0, while desirable values at or beyond a threshold yield an individual desirability of one. Large values of r_1 and r_2 are selected if desirability d_i does not increase substantially until y_i gets significantly close to the target value, T. On the other hand, small values of r_1 and r_2 such as 0.1 indicate that there is a larger acceptable region about T.

After obtaining individual desirability scores for each of the variables of interest, the overall desirability score is obtained by Equation B.11 of Harrington's method. Desirability functions of the LTB, STB, and NTB types are shown in Figure 3.3.

In practice it may be problematic to define shape parameters and weights within their desirability scores. Subjectivity in weight and scale specification is one criticism of desirability scores, however, this subjectivity also exists in the creation of other composite scores. Using desirability functions, the validity of weights and scales can be supported by using a consensus of expert opinion. Furthermore, recent developments have proposed systematic methods of defining shape and weight parameters. This is discussed in detail in Section 3.8.

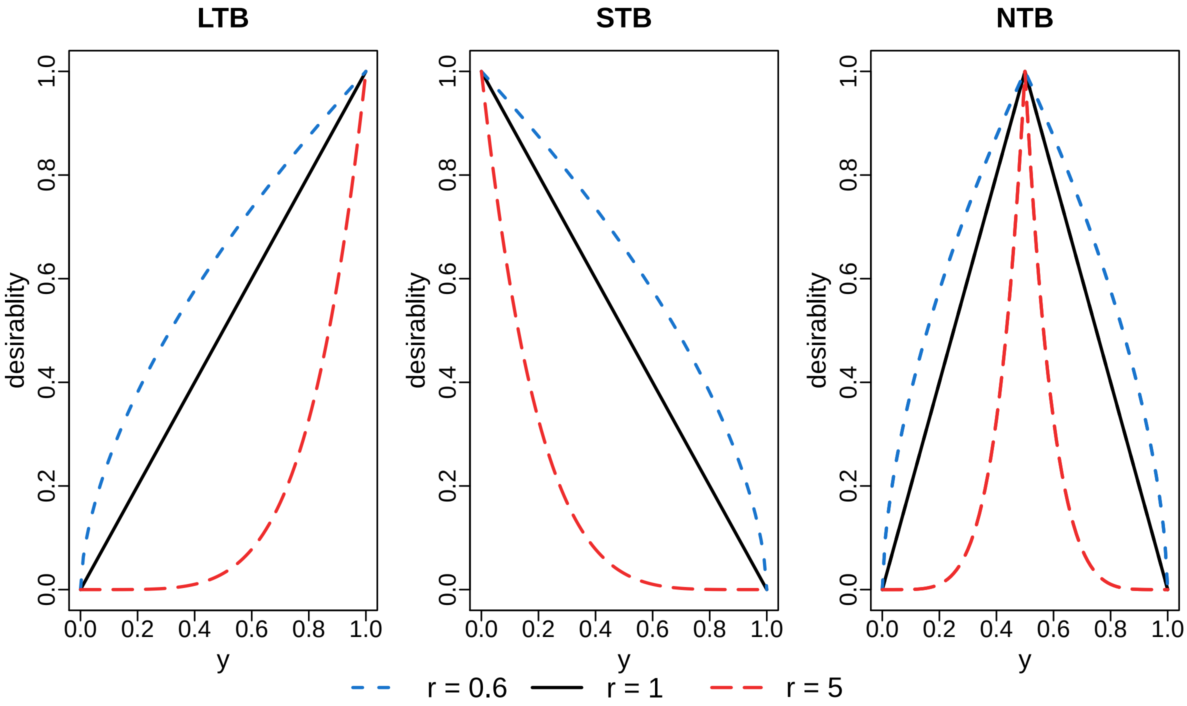


Figure 3.3: Derringer desirability curves with varying shape parameters for L = 0, U = 1, and T = 0.5.

3.1.3 Differentiable Desirability Functions (Castillo et al., 1996)

Castillo et al. (1996) suggested further modifications and presented differentiable desirability functions [25]. His proposed functions were differentiable even without the specification of shape parameters, allowing for efficient gradient-based optimization algorithms, rather than traditional search methods used for Derringer's construction.

Let $f(y)$ be a quartic polynomial that solves conditions elaborated upon in the original paper. Castillo et al. proposed the following desirability function when there is a single non-differentiable point:

$$d_i = \begin{cases} a_1 + b_1 y_i & \text{if } L < y_i \leq T - dy \\ f(y) & \text{if } T - dy < y_i \leq T + dy \\ a_2 + b_2 y_i & \text{if } T + dy < y_i \leq U \\ 0 & \text{otherwise} \end{cases}$$

where dy is a small neighborhood around the non-differentiable point, and a_1, b_1, a_2, b_2 are constants. The rationale was that this modified desirability approach is easier to understand and more intuitive than the shape parameters r, r_1, and r_2.

Weights can still be used in Castillo's method: if the maximum height of all individual desirability functions is set to one at the target of all responses, all the responses are weighted equally; on the other hand, if the target for one response is mapped to a desirability of 0.5, and the target for a second response is mapped to desirability of one, the second response is weighted twice as important as the first.

Castillo et al. applied this modified approach to a wire bonding process from a semiconductor manufacturing process. The drawback of Castillo's approach is that it is only suitable for NTB-type variables.

3.1.4 A Maximin Method (Kim et al., 2000)

Kim et al. (2000) presented a modified desirability score suitable for cases when the overall minimal level of satisfaction with respect to all responses needs to be maximized [53]. The authors' use of a desirability function in exponential form allowed the function to take on various shapes.

Let z represent the distance of an estimated response from its target, standardized by dividing by the maximum allowable deviation. This z parameter then ranges between 0 and 1 for LTB and STB responses, and between -1 and 1 for NTB responses; specifically, we have

For Larger-the-Better (LTB): $z = \frac{y_i - T}{U - T}$
For Smaller-the-Better (STB): $z = \frac{y_i - L}{U - L}$
For Nominal-the-Better (NTB): $z = \frac{U - y_i}{U - L}$.

The individual desirabilities take on an exponential form and are defined by

$$d = \begin{cases} \frac{exp(c) - exp(c|z|)}{exp(c) - 1}, & t \neq 0 \\ 1 - |z|, & t = 0 \end{cases},$$

where c is a constant parameter that allows the model to consider the level of responses' predictive ability. An estimated response with lower predictive ability should have a smaller effect on the desirability score. To incorporate predictive ability of the response, the authors propose replacing c with $c' = c+(1-R^2)(c^max-c)$, where R^2 is the coefficient of determination, and $cmax$ is set large enough so that $d(z)$ with $c = cmax$ is concave such that $d(z)$ has virtually no effect on the optimization process. Other predictive ability measures such as Akaike's information criterion (AIC) can also be used.

The overall desirability score is then $D_{Kim} = maxmin(d_1, d_2, \cdot, d_m)$. The advantages of Kim et al.'s approach is that it does not make any assumptions regarding the form or degree of the estimated response models, and it is robust to potential dependencies among response variables. A drawback of this maximin approach is that it takes the least-worst worse-case-scenario: for example, it would choose a solution with individual desirability levels (0.6, 0.6, 0.6, 0.6) with an overall desirability of $D_{Kim} = 0.6$, over (0.9, 0.9, 0.9, 0.58) with an overall desirability of $D_{Kim} = 0.58$. While this can be a drawback in many scenarios, Kim et al. present cases for which such a strategy is helpful.

3.1.5 Method that Minimizes Average Distance from Targeted Desirabilities (Ch'ng et al., 2005)

Ch'ng et al. (2005) introduced a method to find optimal variable settings assuming normality and homogeneity of error variances [20]. The individual desirabilities are defined as

$$d_i = \frac{(2y_i - (U + L))}{(U - L)} + 1,$$

with $0 \leq d_i \leq 2$. Notice this is consistent with Harrington's individual desirability function plus one. The authors defined a composite score DCh'ng as the weighted average of the distance between an individual desirability score and its value at that characteristic's target value:

$$D_{Ch'ng} = \frac{(\sum_{i=1}^{m} w_i |d_i - d_i(T_i)|)}{m},$$

where $d_i(T_i)$ is the value of the individual desirability function at target value T_i. Since this composite score is a measure of how far desirability scores are from their full potential values, the optimization criterion is to minimize $D_{Ch'ng}$.

The advantages of Ch'ng's method are its interpretability and its minimal subjective specification of parameters; in fact, the number of user-specified parameters (weights) is just the number of responses being evaluated.

3.2 Applications of Desirability Functions in the Medical Field

Medical studies are replete with examples where a composite outcome is used to gauge patient improvement. Desirability functions have only recently begun to show their value in medical settings, since a way of understanding multiple outcomes and how they affect an overall status of a process or a patient can likely result in more effective decisions regarding design or patient care. Interpretation of overall desirability score D and understanding how that relates to patient improvement may vary for each problem or disease. The interpretation needs validation using real data from several trials with a similar cohort of patients.

The first examples shown below use simple shape parameters of one and equal weights when finding the overall desirability, while the later examples are presented in order of increasing complexity.

3.2.1 Bone Drilling Optimization (Pandey et al., 2015)

A common concern during bone drilling is the thermal and mechanical damage inflicted on the bone. In an optimization of a bone drilling process, Pandey et al. (2015) used desirability functions to find the optimal factor levels for two parameters – feed rate and spindle speed – that would result in the optimal combination of two outcomes – temperature and force – during bone drilling in orthopedic surgery [64]. Three levels for each of the two parameters were evaluated, leading to a 32 full factorial design. Shape parameters were set to one and outcome weights were each set to half in a simple application of the desirability function. ANOVA and F-tests were used to determine percentage contribution of both feed rate and spindle speed on composite desirability, with higher F-values indicating stronger influence on the overall desirability.

The authors concluded that feed rate had the highest impact on the outcomes of interest. The factor settings that resulted in the highest overall desirability score (feed rate of 40mm/min and spindle speed of 50rpm) were recommended to minimize temperature and thrust force during bone drilling.

3.2.2 Combining Metabolic Stability & Functionality of Peptides (Van Dorpe et al., 2011)

The evaluation of penetration of biological barriers often examine the transport of drugs through the blood-brain barrier (BBB), which is essential to target brain receptors during the diagnosis or treatment of central nervous system (CNS) disease. Separating blood from the brain, the BBB plays an important role in both allowing beneficial compounds in and harmful compounds out. The metabolic stability of peptides in both plasma and brain tissue is also important because it determines the duration that peptides are presented to the brain.

Van Dorpe et al. (2011) noted that CNS peptides have pharmaceutical potential if they can penetrate the BBB and resist enzymatic degradation for longer periods of time. The authors sought to use desirability functions to select the optimum peptide, with 'optimum' being defined as a compromise between the multiple objectives of BBB transport and metabolic stability [86].

In the determining of drugability of eight different peptides, four different responses were evaluated: blood brain barrier (BBB) influx and efflux, and metabolic stability in brain and in plasma. The four responses were assigned their individual desirability scores via linear desirability functions (r=1), and then combined with equal weights to an overall desirability score via a geometric mean. Derporphin had the highest overall desirability score and thus claimed the highest BBB drugability.

3.2.3 Evaluating Cirhossis (Gennings et al., 2010)

Gennings et al. (2010) observed there was no consistent approach to evaluating cirhossis patients' disease status, and utilized desirability functions to obtain a numerical "wellness" score they named the Relative Wellness Index (RWIc) [33]. Ten experts collaboratively selected ten responses and their respective shape parameters to calculate the overall desirability scores of 109 subjects from the North American Study for the Treatment of Refractory Ascites (NASTRA) dataset.

While former studies of the original NASTRA dataset reported insignificant differences in risk of death or transplant between the two study arms, the authors studied hazard using a Cox PH model and found that a drop of 0.1 in the desirability score (RWIc) was associated with a 21% increase in risk of death or transplant, highlighting the information desirability scores can hold. When dichotomizing the desirability score to ¿ 0.5 and ¡ 0.5, authors identified a significant difference between the two groups.

The composite score was then independently validated using 1342 subjects by constructing the desirability functions and Cox PH model the same way. The validation confirmed that the desirability score was able to predict transplant-free survival. A sensitivity analysis shifting individual desirability functions' shape parameters by ± 10-20% showed robustness of the significant association between transplant-free survival and the overall desirability function, highlighting the ability of desirability functions to reflect more information.

Suleman (2014) and Lazic (2015) also utilized expert opinion by mapping response values to individual desirabilities for rating drug quality for the treatment of intestinal parasites in Ethiopia, and during gene selection and ranking for a breast cancer study, respectively [78, 57]. Both authors noted that traditional methods of manual selection, looking to prior literature, and categorizing candidates as successes or failures due to a single attribute could result in missing potentially interesting and clinically significant candidates that satisfied a spectrum of acceptable attribute combinations.

The example in the next section shows another method to utilize professional evaluation in the building of individual desirability functions.

3.2.4 Assessing Scleroderma Progression (Wong et al., 2007)

Wong et al. (2007) utilized desirability functions to assess the progression of scleroderma disease in a survival analysis setting [91]. Kaplan-Meier plots for single outcome variables and multiple endpoints helped the authors decide that patients with a 10% drop in lung performance tests should be monitored. The plots further established the importance of certain outcomes - skin score, modified health assessment questionnaire (mHAQ), patient global assessment, physician global assessment (pga), and FVCP - when assessing the overall health of a scleroderma patient.

Physician global score ranged from -2 to +3, and was used as a calibrating variable: patients were broken into buckets per their physician global assessment score at the one-year time point, and each bucket was assigned a desirability value. Specifically, pga scores of -2, -1, 0, 1, 2, and 3 had initial desirabilities of 0, 0.2, 0.4, 0.6, 0.8, and 1 respectively. Percentage mean and median change for skin score and mHAQ were calculated for each bucket.

Desirability functions were then plotted, with desirability score on the y- axis and the mean or median percent change on the x- axis. For example, the bucket of patients with pga of 0 were assigned desirability of 0.4, whilst the same bucket's median skin score percent change was -7%. Meanwhile, the bucket of patients with pga of three were assigned desirability one, whilst the same bucket's median skin score percent change was -23%. Each of these median percent change scores for each bucket was plotted against the desirability score previously defined. In this way, the individual desirability functions were plotted, and intermediate values could then be mapped to their corresponding individual desirabilities.

The authors then proceeded to show how the individual desirabilities could be combined with different weights to obtain the overall desirability score. While a scaling parameter was not specifically defined here, using median percent change within each plausible value of physician global score thereby defined the shape of each individual desirability curve.

3.2.5 Clinical Trials Sample Size (Fransen et al., 2009)

Fransen et al (2009) observed that there is no one agreed upon measure to assess progression of psoriatic arthritis, and noted that combining outcomes in chronic inflammatory rheumatic diseases, such as rheumatoid arthritis and systemic sclerosis, could lead to increased precision and validity. This is because randomized controlled trials studying these diseases often use binary or dichotomized responses as the primary outcome, which generally lead to loss in power relative to scenarios using continuous measures. The authors recalculated the required sample size of a previous clinical trial had desirability scores been used, and concluded that the method utilizing desirability scores had higher efficiency [31].

Desirability functions combined four key responses, incorporating functional disability as well as the patient's self-evaluation of disease progress, into an overall composite score that the authors named CRISS. The selection of these four responses was to be consistent with the responses that construct the Disease Activity Score (DAS28), one of the common assessment scores used in rheumatoid arthritis. The authors retrospectively used the medians of 44 expert ratings for boundaries of clinical states of remission and low, moderate, and high arthritic activity. Remission was equated to Harrington's "acceptable and excellent", low disease activity to "acceptable and good", moderate disease activity to "poor/borderline", and high disease activity to "unacceptable".

In a Phase II trial seeking to evaluate the efficacy of a TNF-inhibitor as a therapy for psoriatic arthritis, the treatment was found to be statistically more effective than placebo when evaluating group differences via t-test in scores at 16 weeks. The authors calculated the effect size and the relative efficiency to compare the efficiency of the individual desirability scores with that of the original variables. The number of patients required for a specified statistical power is inversely proportional to the squared effect size. The authors showed that in the TNF-inhibitor trial, the individual desirability scores had higher efficiency compared to when using the original outcomes. For example, when using the individual desirability function for tender joint count to calculate sample size, only 41% of the sample size was needed as compared to when using the clinical score itself.

The authors observed that a meaningful approach would be to take the minimum score of a set of scores (as in Section 2.4) to show that a treatment does better on all components of arthritic activity. However, for the sake of efficiency, a simple geometric mean was taken to combine the four individual desirability scores into an overall desirability score. Compared to sample size based on DAS28, the sample size based on the authors' new overall desirability score was 13% smaller. Furthermore, the authors showed the robustness of their results by performing a sensitivity analysis with different percentiles of the expert ratings defining the individual desirability curves. Their work is an excellent example of how desirability functions can be used effectively in biomedical research.

3.2.6 Discrete & Continuous Outcomes in a Dose-Response Study (Coffey et al., 2007)

Coffey et al. (2007) discussed the advantages of using desirability functions to analyze outcomes in a dose-response study [21]. One unique aspect of their paper is the use of desirability scores as an outcome in further modeling. The neurotoxicity of a five-pesticide mixture was assessed using one control dose and six experimental fixed doses. Five endpoints were assessed in the examination of toxicity: motor activity, gait abnormality, tail-pinch response, and the neurochemical endpoints of cholinesterase activity in the brain and in whole blood.

Five expert toxicologists responded to a questionnaire which asked the respondents to characterize the level of toxicity associated with various responses. These responses were not just continuous, but also included binary and ordinal endpoints. For example, for the outcome gait score, the five possible responses were ordinal: “none”, “slightly abnormal”, “somewhat abnormal”, “markedly abnormal”, and “severely abnormal”. The expert toxicologist respondents then drew a line on a continuum between absolutely no toxicity and the most severe toxicity, indicating their belief of how toxic the drug was based on response levels. The distance between the left boundary of no toxicity and the respondent’s marked line was converted to a proportion which was used as the individual desirability score (e.g. a line drawn halfway between no toxicity and most severe toxicity would be given a score of 0.5).

Desirability functions were then modeled using a simple logistic cumulative distribution function (CDF) for the strictly increasing endpoint of gait normality, and two logistic CDFs for the remaining concave-shaped responses (motor activity, cholinesterase activity in the brain and in the blood, and tail-pinch response).

Individual desirability scores were combined into an overall desirability score using the geometric mean with equal weighting. The data suggested a nonlinear dose-response with a dose threshold, so the authors used a nonlinear exponential model with a threshold parameter, using desirability score as the outcome. A summary threshold for all outcomes was obtained, and is interpreted as an estimate of the smallest dose where evidence of toxicity is present in at least one of the endpoints. A sensitivity analysis using different desirability functions yielded similar results, showing the robustness of the authors’ methods.

3.2.7 Using Desirability Functions for Vaccine Formulation (Dewe et al., 2016)

In vaccine formulation, it is important to elicit a strong immune response while maintaining reactogenicity and safety profile. Regulators have increasingly demanded justification of inclusion and doses of different vaccine components. While recent literature showed vaccine formulation was still heavily based off of descriptive studies, Dewé et al. (2016) used desirability functions for the first time in the formulation of a vaccine [27]. Furthermore, the authors obtained confidence intervals to quantify uncertainty associated with desirability scores using the bootstrap method.

The primary immunogenicity endpoint was hemagglutination inhibition (HI) titer 21 days after vaccination, a measure of antibodies circulating in the body. The geometric mean titer (GMT) was defined as the anti-log of a log HI mean estimate, while the geometric mean ratio (GMR) was defined as the ratio of two groups’ GMTs. The reactogenicity endpoint was frequency of severe solicited general adverse events (AEs) observed within seven days after vaccination.

Desirability functions are beneficial in vaccination component selection because the effects of multiple components (MPL and AS03, in this case) as well as their interaction effects on multiple outcomes (immunogenicity and reactogenicity) can be evaluated. The effects of the two components on HI titer were evaluated in a linear regression model. The difference between the least square mean and the arithmetic mean of the

HI titers was used to estimate the GMR. The effects of the two components on the reactogenicity endpoint were evaluated in a logistic regression model. The authors chose a logistic curve to shape their desirability functions.

The lower 90% confidence interval of the predicted HI titer less the average HI titer was used in the logistic desirability function to calculate three sub-desirability scores for immunogenicity, one for each of three vaccine strains. The immunogenicity desirability score was then calculated by taking the geometric mean of the three strains' individual immunogenicity sub-desirabilities. The probability of experiencing an adverse event as determined by the reactogenicity logistic regression model was used in an exponential function to calculate a desirability score for reactogenicity.

Although the authors note the challenge of weight selection, they prioritize immunogenicity improvement, giving the immunogenicity desirability score a weight of 0.6, and the reactogenicity desirability score a weight of 0.4. The overall desirability score was thus able to characterize the desirability of any formulation in the experimental domain. A sensitivity analysis showed that rankings varied slightly depending on weight choices. To account for variance stemming from the linear model for immunogenicity and the logistic model for reactogenicity, the authors employed a bootstrap method, noting that the distribution of the ranks do not alter the ranking as decided by the desirability score in the original analysis.

The authors note that the desirability functions allowed for a broader range of possible solutions to be explored while remaining clinically executable in the medical field, since fewer subjects were enrolled than would have been necessary in a formal clinical trial.

3.2.8 Desirability Scores of Randomization Sequences (Schindler, 2016)

Schindler (2016) utilized desirability scores to assess the desirability of various randomization sequences, where a randomization sequence is the order in which treatment arms are assigned to enrolling clinical trial subjects [75]. He focuses on two types of biases in randomized clinical trials: the first is selection bias, where an investigator finds a way to intervene with treatment assignment to patients; the second is chronological bias, where treatment quality could change throughout the course of a trial, whether more co-medications become available or a physician's skill improves. The selection bias can be described as an investigator's ability to guess the next allocation in a balanced trial design depending on how many patients are in the experimental arm and how many are in the control arm. The two biases are then able to be summarized in a single score that assesses the desirability of a design.

In a simple example, Schindler evaluates a randomization sequence using the proportion of correct guesses, and the Type 1 and Type II errors in the presence of chronological bias. In the overall desirability function, half the weight is placed on selection bias (proportion of correct guesses), and the other half on chronological bias (Type I and Type 2 errors).

The overall desirability of a design can then be assessed by calculating the desirability of each possible randomization sequence resulting from that design, and then looking at the expected overall desirability, the standard deviation of the expected overall desirability, and the probability that the expected overall desirability is zero.

3.2.9 Systematic Definition of Shape & Weight Parameters (Chen et al., 2015)

Shape and weight parameters are important determinants of the desirability score. The subjective definitions of shape and weight parameters may have contributed to the biomedical field's relatively slower adoption of desirability functions, since - in the case of health data - physicians or experts may not agree on how different outcomes correspond to different degrees of disease progression, and furthermore, the selection of physicians or experts who define the impact of different outcomes on disease progression are subject to selection bias. While entropy weight theory is a useful method to obtain objective weights by measuring overall variability of responses [77], it is often not applicable in biomedical research because its use is limited to quality (categorical) characteristics. Zhang et al. (2013) and Zhou et al. (2015) show how entropy weights can be used in the construction of desirability functions to assess colloidal gas alphorn [95, 97].

Chen et al., proposed a systematic method of defining shape parameters and weights that was applicable also to continuous outcomes [18]. Since an individual desirability score should approximate a gold standard as close as possible, Chen et al. noted that we can replace d_i in our previous Equations B.8 - B.10 with d_g, a given gold standard (e.g. physician global assessment score) rescaled between 0 and 1. For example, for STB variables:

$$d_g = (\frac{U - y_i}{U - L})^{r_i} + \epsilon_i,$$

where ϵ_i is an approximation error. Nonlinear least squares can then be used to estimate shape parameter r_i. Since the gold standard was used as the outcome of the nonlinear least squares model, the resulting shape parameter then places the desirability curve of the response into perspective relative to the gold standard. The estimated r_i can then be used in turn to obtain the individual desirability score di using the original STB equation in Equation B.9. The same method can also be applied to LTB and NTB variable types.

After obtaining the individual desirability scores for all outcomes, weights can then be found in a similar manner, using

$$d_g = (d_1^{w_1} d_2^{w_2} ... d_m^{w_m})^{\frac{1}{\sum_{i=1}^{m} w_i}} + \epsilon.$$

An arbitrary component's weight can be set to one. Then, nonlinear least squares can again be used to solve for the remaining weights, with the constraint that they be non-negative. After estimation of the weights, they can be rescaled such that they sum to one. After the weights are calculated, the overall desirability score can be constructed as before using Equation B.11. The methodology presented by Chen et al. provides a strong methodology for adoption in future biomedical research.

3.2.10 Desirability Functions in a Longitudinal Setting (Chen et al., 2015)

Chen et al. (2015) also sought to use desirability scores in a longitudinal study, and thus introduced a new method of incorporating desirability scores over time [18]. If an overall desirability score at time t is denoted by D_t, then a modified L_p norm can be used to combine these overall desirability scores at different time points into a single score. If there are t time points, then a composite overall desirability score can be constructed:

$$D* = [\frac{\lambda_1 D_{t_1}^p + ... + \lambda_k D_{t_k}^p}{\lambda_1 + ... + \lambda_k}]^{1/p},$$

where λ_j is the weight of time point t_j, j = 1,. . . ,k. The advantage of this approach is that when one variable may not be measured as consistently as others, its weight at that time point can be zero, while still utilizing the information it provides for the other time points.

The authors utilized the new method of obtaining scale and weight parameters and the new incorporation of desirability scores from different time points to compare the two arms of a clinical trial of 168 scleroderma patients. These patients were treated with placebo or oral type I collagen over the course of 12 months, and followed up at month 15. Following clinical practice, patient's global assessment of health (pga) and physician's overall assessment of disease activity (poa) were used as gold standards during estimation of shape parameters and weights. Shape and weight estimation varied significantly depending on the choice of gold standard. The lower bound and upper bounds of each variable were specified as observed ranges from baseline to month 12.

When patient's global assessment of health was used as the gold standard, many of the components had higher desirabilities. Although the choice of the gold standard affects the shape parameters and the weight estimates, the overall interpretation of disease progression was consistent across the results, finding that the scleroderma patients in this trial failed to show a significant progression over time.

3.3 Innovations in Desirability Functions

Desirability functions have started to emerge in the biomedical field, but their applications are often simplified. Perhaps one reason it has not yet been widely adopted in the medical field is that its traditional version does not address variance or uncertainty. A handful of authors have addressed this issue with newly proposed methods: Wu (2004) defined a new individual desirability function that incorporates variation and correlation [93, 94]; Govaerts et al. (2005) maximized the expected overall desirability and introduced theoretical confidence intervals for overall desirabilities [35]; Monteagudo (2008) introduced an overall desirability determination coefficient to assess the quality of the measure [23]; He et al. (2012) proposed to maximize the minimum desirability score of confidence regions rather than single prediction values [42]; and Chen et al. (2012) introduced a secondary individual desirability function that accounts for the variance of the predicted response [17]. Despite the importance of accounting for variability in the health field, the methods of Wu, Govaerts, and He have yet to be implemented in biomedical research. Their methods are detailed in this section.

3.3.1 Incorporation of Variation & Correlation of Responses into Desirability Functions (Wu et al., 2000; Wu, 2004)

Wu et al. (2000) noted that a disadvantage of the traditional desirability function is its inability to consider variation and correlation between responses [92]. While Taguchi's robust design method reduces variation by focusing on a single quality characteristic to optimize the other parameters [80], finding the optimal settings for all variables is difficult. Indeed, improving the values of one variable may come hand in hand with the digression of another, especially amongst negatively correlated variables. A modified double-exponential desirability function was proposed.

Let c, c_1, c_2 be scalar constants, and let r, r_1, and r_2 be shape parameters for the desirability function. Wu and Hamada's double-exponential desirability functions are defined as:

For Larger the Better (LTB):

$$d_i = \begin{cases} \frac{1-exp(-cy_i^r)}{exp(-cL^r)} & \text{for } L \leq y_i < \infty \\ 0 & \text{for } y_i < L \end{cases}$$

For Smaller the Better (STB):

$$d_i = exp(-c|y_i - L|^r), \text{for } L \leq y_i < \infty$$

For Nominal The Better (NTB):

$$d_i = \begin{cases} exp(-c_1|y_i - T|^{r_1}) & \text{for } -\infty < y_i \leq T \\ exp(-c_2|y_i - T|^{r_2}) & \text{for } T \leq y_i < \infty. \end{cases}$$

One drawback of this approach is that it requires more user-specified parameters than the original Derringer-Suich desirability function. The rest of the overall desirability score D is constructed as in Equation B.11.

Wu (2004) builds on this by incorporating a loss function of the prediction function of y, which can be represented as $L(y(x))$, and can further be expanded in a Taylor series about target T of y [9]. The loss function is inspired by Taguchi's definition of quadratic quality loss functions. Let k_j be the loss coefficient of characteristic or response y_i when it deviates from the target T_i. Let k_{ij} be the correlated loss coefficient of quality characteristics y_i and y_j when they simultaneously deviate from their respective targets, T_i and T_j. Let ρ_{ij} be the correlation coefficient between y_i and y_j, so that the quality loss can be approximated by a quadratic function as:

$$\begin{aligned} Loss &= \sum_{i=1}^{m} k_i(y_i - T_i)^2 + \sum_{i=1}^{m}\sum_{i<j}^{m} k_{ij}(y_i - T_i)(y_j - T_j) \\ &= \sum_{i=1}^{m} [\tilde{k}_i \cdot ((\hat{y}_i - T_i)^2 + \hat{\sigma}_i^2] + \sum_{i=1}^{m}\sum_{i<j}^{m} \tilde{k}_{ij} \cdot ((\hat{\rho}_{ij}\hat{\sigma}_i\hat{\sigma}_j + (\hat{y}_i - T_i)(\hat{y}_j - T_j)) \\ &:= \sum_{i=1}^{m} Loss_i + \sum_{i=1}^{m}\sum_{i<j}^{m} Loss_{ij}. \end{aligned}$$

The double-exponential desirability value for y_i is then defined as

$$d_{ii} = exp[-Loss_i],$$

and the correlated desirability value between y_i and y_j is

$$d_{ij} = \begin{cases} 1 & Loss_{ij} < 0 \\ exp[-Loss_{ij}] & Loss_{ij} > 0. \end{cases}$$

With m total responses, the total overall desirability D_{WU} is then

$$D_{WU} = [\prod_{i=1}^{m} \prod_{i<j}^{m} d_{ii} d_{ij}]^{1/m}.$$

Wu utilized this method to identify the ideal factor settings for 8 variables for a plasma-enhanced chemical vapor deposition process where deposition thickness and refractive index were both outcomes of interest, as well as a second case to identify the ideal settings for 6 variables for a polysilicon deposition process where surface defects, thickness, and deposition rate were outcomes of interest. Wu compared his new desirability scores with those derived from factor settings identified in previous literature, showing a significant increase in total desirability. Wu's desirability function has yet to be utilized in health outcomes' research.

3.3.2 Distribution of the Desirability Score (Trautmann, 2004, 2005)

Trautmann (2005) acknowledges that Harrington's desirability function and Derringer-Suich's modifications are widely accepted (e.g. Tyuev et al., 1997; Wu et al., 2000; Averill et al., 2001; Ben-Gal et al., 2002), yet observes that confidence intervals are not easily accessible due to a lack of understanding of the distribution of the desirability score [85].

Trautmann derives the distribution of Harrington's desirability functions with the assumption that the components evaluated are normally distributed. Let component Y_i $N(\mu_i, \sigma_i^2)$. Then, for one-sided variables:

$$Z_i := exp[-Y_i'] \ Lognormal(\tilde{\mu}_i, \tilde{\sigma}_i^2),$$

with $\tilde{\mu}_i = -(b_{0i} + b_{1i}\mu_i)$ and $\tilde{\sigma}_i^2) = (b_{1i})^2$. Similarly, for two-sided variables:

$$X_i := |Y_i'| \ FoldedNormal(\tilde{\mu}_i, \tilde{\sigma}_i^2),$$

with $\tilde{\mu}_i = \frac{2}{U_i - L_i}\mu_i - \frac{U_i + L_i}{U_i - L_i}$ and $\tilde{\sigma}_i^2 = (\frac{2}{U_i - L_i})^2 \sigma_i^2$. Then, by the density transformation theorem, the density and distribution functions are:

For one-sided variables:

$$f_{D_i}(d_i) = -\frac{1}{\sqrt{2\pi}\tilde{\sigma}_i log(d_i) d_i} \cdot exp[-\frac{1}{2\tilde{\sigma}_i^2}(log(-log(d_i)) - \tilde{\mu}_i)^2]$$

$$F_{D_i}(d_i) = 1 - \Phi[(log(-log(d_i)) - \tilde{\mu}_i)/\tilde{\sigma}_i].$$

For two-sided variables:

$$f_{D_i}(d_i) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}_i d_i n_i} \cdot (-log(d_i))^{1/n_i - 1}$$
$$[exp(-((-log(d_i))^{1/n_i} - \tilde{\mu}_i)^2/2\tilde{\sigma}_i^2) + exp(-((-log(d_i))^{1/n_i} + \tilde{\mu}_i)^2/2\tilde{\sigma}_i^2)]$$

$$F_{D_i}(d_i) = 2 - \Phi\Bigg[\frac{((-log(d_i))^{1/n_i} - \tilde{\mu}_i)}{\tilde{\sigma}_i}\Bigg] - \Phi\Bigg[\frac{((-log(d_i))^{1/n_i} + \tilde{\mu}_i)}{\tilde{\sigma}_i}\Bigg],$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal.

The distribution of the overall desirability score when using the geometric mean is analyzed separately for one- and two- sided type variables. The first step involves rewriting the overall desirability score D:

For one-sided variables, the sum of lognormal random variables is involved:

$$D := (\prod_{i=1}^{m} d_i)^{1/m} = (\prod_{i=1}^{m} exp[-exp[-Y_i']])^{1/m} = (exp[-\sum_{i=1}^{m} exp[-Y_i']])^{1/m}.$$

For two-sided variables:

$$D := (\prod_{i=1}^{m} d_i)^{1/m} = (\prod_{i=1}^{m} exp[-|Y_i'|^{n_i}])^{1/m} = (exp[-\sum_{i=1}^{m} |Y_i'|^{n_i}])^{1/m}.$$

Then, given m one-sided variables with individual desirabilities d_i, the density and distribution function of the overall desirability score D when combined via geometric mean can be approximated. Since $D := (exp[-\sum_{i=1}^{m} exp[-Y_i']])^{1/m}$ and the exponent $-\sum_{i=1}^{m} exp[-Y_i'] \sim Lognormal(\mu^*, \sigma^{*2})$, then $Z := exp[-exp[-Y_i']]$ approximately follows a double lognormal distribution $DLN(\mu^*, \sigma^{*2})$. Then, since $D = Z^{1/m}$, the approximate distribution of a one-sided overall desirability score is:

$$f_d(D) \approx -\frac{1}{\sqrt{2\pi}\sigma^* log(D) D} exp\big[-\frac{1}{2\sigma^{*2}}(log(-m \cdot log(D)) - \mu^*)^2\big]$$

$$F_D(D) \approx 1 - \Phi\big[\frac{log(m) + log(-log(D)) - \mu^*}{\sigma^*}\big].$$

On the other hand, given two independent two-sided variables $Y_i (i = 1, 2)$ with individual desirability scores d_i and kurtosis parameter $n_i = 1 \forall i$, the overall desirability score has the density function:

$$\begin{aligned} f_D(D) = & \frac{\sqrt{2}}{2D\sqrt{\pi(\tilde{\sigma}_2^2+\tilde{\sigma}_1^2)}} \cdot \\ & \Bigg(exp[-\frac{(-2log(D)-\tilde{\mu}_1-\tilde{\mu}_2)^2}{2(\tilde{\sigma}_2^2+\tilde{\sigma}_1^2)}] \times erf(\frac{((-2log(D))\tilde{\sigma}_2^2-\tilde{\mu}_1\tilde{\sigma}_2^2+\tilde{\mu}_2\tilde{\sigma}_1^2)}{\tilde{\sigma}_2\tilde{\sigma}_1\sqrt{2}\sqrt{\tilde{\sigma}_2^2+\tilde{\sigma}_1^2}}) \\ & (exp[-\frac{(-2log(D)-\tilde{\mu}_1-\tilde{\mu}_2)^2}{2(\tilde{\sigma}_2^2+\tilde{\sigma}_1^2)}] \times erf(\frac{((-2log(D))\tilde{\sigma}_2^2-\tilde{\mu}_1\tilde{\sigma}_2^2+\tilde{\mu}_2\tilde{\sigma}_1^2)}{\tilde{\sigma}_2\tilde{\sigma}_1\sqrt{2}\sqrt{\tilde{\sigma}_2^2+\tilde{\sigma}_1^2}}) \\ & (exp[-\frac{(-2log(D)-\tilde{\mu}_1-\tilde{\mu}_2)^2}{2(\tilde{\sigma}_2^2+\tilde{\sigma}_1^2)}] \times erf(\frac{((-2log(D))\tilde{\sigma}_2^2-\tilde{\mu}_1\tilde{\sigma}_2^2+\tilde{\mu}_2\tilde{\sigma}_1^2)}{\tilde{\sigma}_2\tilde{\sigma}_1\sqrt{2}\sqrt{\tilde{\sigma}_2^2+\tilde{\sigma}_1^2}}) \\ & (exp[-\frac{(-2log(D)-\tilde{\mu}_1-\tilde{\mu}_2)^2}{2(\tilde{\sigma}_2^2+\tilde{\sigma}_1^2)}] \times erf(\frac{((-2log(D))\tilde{\sigma}_2^2-\tilde{\mu}_1\tilde{\sigma}_2^2+\tilde{\mu}_2\tilde{\sigma}_1^2)}{\tilde{\sigma}_2\tilde{\sigma}_1\sqrt{2}\sqrt{\tilde{\sigma}_2^2+\tilde{\sigma}_1^2}}) \\ & (exp[-\frac{(-2log(D)-\tilde{\mu}_1-\tilde{\mu}_2)^2}{2(\tilde{\sigma}_2^2+\tilde{\sigma}_1^2)}] \times erf(\frac{((-2log(D))\tilde{\sigma}_2^2-\tilde{\mu}_1\tilde{\sigma}_2^2+\tilde{\mu}_2\tilde{\sigma}_1^2)}{\tilde{\sigma}_2\tilde{\sigma}_1\sqrt{2}\sqrt{\tilde{\sigma}_2^2+\tilde{\sigma}_1^2}}) \\ & (exp[-\frac{(-2log(D)-\tilde{\mu}_1-\tilde{\mu}_2)^2}{2(\tilde{\sigma}_2^2+\tilde{\sigma}_1^2)}] \times erf(\frac{((-2log(D))\tilde{\sigma}_2^2-\tilde{\mu}_1\tilde{\sigma}_2^2+\tilde{\mu}_2\tilde{\sigma}_1^2)}{\tilde{\sigma}_2\tilde{\sigma}_1\sqrt{2}\sqrt{\tilde{\sigma}_2^2+\tilde{\sigma}_1^2}}) \\ & (exp[-\frac{(-2log(D)-\tilde{\mu}_1-\tilde{\mu}_2)^2}{2(\tilde{\sigma}_2^2+\tilde{\sigma}_1^2)}] \times erf(\frac{((-2log(D))\tilde{\sigma}_2^2-\tilde{\mu}_1\tilde{\sigma}_2^2+\tilde{\mu}_2\tilde{\sigma}_1^2)}{\tilde{\sigma}_2\tilde{\sigma}_1\sqrt{2}\sqrt{\tilde{\sigma}_2^2+\tilde{\sigma}_1^2}}) \\ & (exp[-\frac{(-2log(D)-\tilde{\mu}_1-\tilde{\mu}_2)^2}{2(\tilde{\sigma}_2^2+\tilde{\sigma}_1^2)}] \times erf(\frac{((-2log(D))\tilde{\sigma}_2^2-\tilde{\mu}_1\tilde{\sigma}_2^2+\tilde{\mu}_2\tilde{\sigma}_1^2)}{\tilde{\sigma}_2\tilde{\sigma}_1\sqrt{2}\sqrt{\tilde{\sigma}_2^2+\tilde{\sigma}_1^2}})\Bigg), \end{aligned}$$

where $erf(x) = 2 \cdot \Phi(\sqrt{2}x) - 1$.

The authors were unable to derive an analytical form of the cumulative distribution function for overall desirability scores consisting of two-sided variables. The densities and distribution functions for Kim's minmax D were also derived in Trautmann's work.

The significance of this work is its contribution towards quantifying uncertainty of both individual desirability scores as well as overall desirability scores. The authors note that extensions to understanding the distributions of desirability scores of the Derringer-Suich construction would be valuable.

3.3.3 Optimizing Expected Desirability and Defining Confidence Intervals (Govaerts et al., 2005)

Govaerts et al. (2005) noted that desirability functions are sensitive to validity of model predictions, and sought to quantify uncertainty stemming from this dependency [35]. Since the desirability score is a random variable, the authors proposed to optimize the expectation of the desirability. To differentiate between the two concepts of maximizing overall desirability versus maximizing the expected overall desirability, we let $D^C(x)$ denote the classic overall desirability score which takes its maximum at $x_{opt}C$, and $D^N(x)$ be the "new" expected overall desirability score which takes its maximum at $x_{opt}N$. Then,

$$D^N(x) = E[\prod_{i=1}^{p}(d_i)^{w_i}] \neq \prod_{i=1}^{p}(d_i(E[Y_i|x]))^{w_i} = D^C(x).$$

The asymptotic distribution of the optimized expected desirability score can be used to build confidence intervals for the true expected desirability score, where we replace unknown quantities with their classic estimators as observed from the data, marked by hats.

Since the average overall desirability score could not be expected to be asymptotically normal due to its range of [0, 1], a logit transformation can be used so that the support ranges $[-\infty, +\infty]$. Then, using the delta method, the confidence interval for $D^N(x)$ is:

$$\frac{exp[logit(\hat{D}^N(x)) \pm z_{1-\alpha/2}\sqrt{\hat{Var}[logit(\hat{D}^N(x))]}}{1 + exp[logit(\hat{D}^N(x)) \pm z_{1-\alpha/2}\sqrt{\hat{Var}[logit(\hat{D}^N(x))]}}.$$

The performance of this theoretical confidence interval obtained by the delta method was evaluated through simulation by comparing the observed 2.5th and 97.5th percentiles of a simulated distribution with the theoretical 95% confidence interval. It was concluded that the approximated asymptotic distribution given by the delta method provides accurate confidence intervals even with small samples. Furthermore, if the distribution of $D(x)$ has no analytical form, a prediction interval for $D(\hat{x}^C_{opt})$ can still be made using the quantiles of an empirical distribution obtained through simulation.

3.3.4 Desirability Determination Coefficient and A Ranking Method (Monteagudo et al., 2008)

Monteagudo et al. (2008) noted the importance of strong predictive models, and introduced four concepts that could be used with desirability functions: an overall desirability's determination coefficient (R^2_D), a leave-one-out cross-validation (LOO-CV) determination coefficient (Q^2_{LOO}), a ranking algorithm, and a ranking quality index Ψ^* [23].

The overall desirability's determination coefficient (R^2_D) is a measure of the effect of the set of variables (X's) in reducing the uncertainty when predicting overall desirability values using linear regression models, and is defined as:

$$R^2_D = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum(D_{y_i} - D_{\hat{y}_i})^2}{\sum(D_{y_i} - \overline{D}_{y_i})^2},$$

where SSE is the Sum Square Error, $SSTO$ is the Total Sum of Squares, and $\overline{D}_{y_i}$ and $D_{\hat{y}_i}$ are the mean value of the overall desirability score for the observed yi responses, and the overall desirability score for the predicted response, respectively. Larger R_D^2 indicates more certainty in predicting D with a set of independent variables X. The analogous leave-one-out cross-validation (LOO-CV) determination coefficient Q_{LOO}^2 is used to establish the reliability of the method in predicting D:

$$Q_D^2 = 1 - \frac{SSE_{LOO-CV}}{SSTO} = 1 - \frac{\sum(D_{y_i} - D_{\hat{y}_i(LOO-CV)})^2}{\sum(D_{y_i} - \overline{D}_{y_i})^2},$$

where SSE_{LOO-CV} is the leave-one-out cross validation square sum of residuals and $D_{\hat{y}_{i(LOO-CV)}}$ is the predicted overall desirability when using LOO-CV.

Monteagudo then proposed a ranking algorithm that uses desirability functions to rank candidates under consideration by determining similarity with the optimal candidate as determined by the highest desirability score. To do so, a Δi is calculated for each candidate i over m characteristics, defined as the sum of weighted Euclidean distances of each X value from its optimal value: $\Delta_i = \sum_{X=1}^{m} \delta_{i,X} \cdot w_x$, where $\delta_{i,X} = |X_i - X_{optimal}|$. Evaluating $\delta_{i,X}$ allows an analyst to see how each variable influences the overall desirability D, and for more degrees of freedom to finding optimal weights.

The STB desirability function is then applied to these Δ_i. Nonlinear least-squares regression is used to solve for optimal set of weights w_x so that the difference between the individual desirabilities of Δ_i and D_i are minimized. This minimization is to ensure that the information represented by the X's and combination of the X's that contribute to the overall desirability are as close as possible. Ranking can be done then with candidates with highest values of individual desirabilities of Δ_i.

A quantitative criterion the authors call a ranking quality index is defined. Let OT be a true order list of candidates ordered by decreasing overall desirability D, labeled $1, \ldots, n$; OW be the worst-order list ordered by increasing overall desirability D; and OR be a list resulting from the ranking algorithm, ordered by increasing Δ_i. Obtain individual desirabilities (STB) for each of the rank values in these three lists, setting $L = 1$ and $U = n$, the number of candidates under consideration. Let d_i^{OT} be the individual desiraiblities list for the true order, d_i^{OW} be the individual desirabilities list for the worst-order, and d_i^{OR} be the individual desirabilities list that result from the ranking algorithm. The quality of the ranking can then be assessed by seeing how far $|d_i^{PT} - d_i^{OR}|$ is from zero. The ranking quality index is defined as:

$$\Psi^* = |\frac{\sum_{i=1}^{n} |d_i^{OT} - d_i^{OR}|}{n}| \cdot \frac{2}{|\frac{\sum_{i=1}^{n} |d_i^{OT} - d_i^{OW}|}{n}|},$$

where the second term allows Ψ^* to range between [0, 1], where a Ψ^* of zero indicates a perfect ranking, and a Ψ^* of one indicates the worst ranking.

Monteagudo applied his proposed methods to rank 95 drug candidates, with antibacterial activity and cytotoxicity as the outcomes of interest with equal importance. His method was an improvement relative to the standard approach of the pharmaceutical industry, which was to optimize objectives one at a time, often leading to less-than-optimal results. Monteagudo's overall desirability function had an R_D^2 of 0.7 and a Q_D^2 of 0.63, indicating good statistical quality of the overall desirability score and an adequate level of reliability on the method used to predict D. The overall desirability was then optimized to obtain the levels of the descriptors included in the linear regression models that would simultaneously produce the most desirable combinations of the properties.

3.3.5 Maximin Desirability Scores of Confidence Regions (He et al., 2012)

He et al. (2012) created a robustness measure for overall desirability, pointing out that a true response $y_i(x)$ is usually unknown, and that models are used to fit each response using empirical data [42].

Let there be k factors $x = (x_1, x_2, \ldots, x_k)$ that influence m independent responses, $y_1, \ldots, y_m$. Assume a response y_i can be approximated with $y_i = f_i(x)^T\beta_i + \epsilon_i$, where ϵ_i $N(0, \sigma^2)$. Consequently, $f_i(x)$ reveals a vector of coefficients for the betas: a first-order linear model is represented by $f_i(x) = (1, x_1, \ldots, x_k)^T$, and a second-order quadratic model is represented by $f_i(x) = (1, x_1, \ldots, x_k, x_1^2, \ldots, x_k^2, x_1x_2, \ldots, x_{k-1}x_k)^T$. Assuming $X_i = (f_i(x_1), \ldots, f_i(x_n))^T$ is full column rank, then the least squares estimate of beta is $\hat{\beta}_i = (X_i^T X_i)^{-1}(X_i)^T y_i$. Since the expected value of the error term is zero, the predicted response at x is then $\hat{y}_i = f_i(x)^T\hat{\beta}_i$, and the variance of the predicted response is $var(\hat{y}_i) = \sigma_i^2 v_i(x)$ where $v_i(x) = f_i(x)^T(X_i^T X_i)^{-1}f_i(x)$. Then the $(1-\alpha)$ confidence interval of the mean response $y(x)$ is given by

$$[LC, UC] = [\hat{y}(x) \pm t_{\alpha/2, n-p}\sqrt{\hat{\sigma}^2 v_i(x)}].$$

If $[LC, UC]$ is the $(1-\alpha)^{1/m}$ confidence interval on a $y_i(x)$ where $i = (1, 2, \ldots, m)$, then a simultaneous joint confidence region for m independent responses with family error rate equal to α is then the product of these intervals: $[LC_1, UC_1]X[LC_2, UC_2]X \ldots X[LC_m, UC_m]$.

If the true responses take any point $\eta = (\eta_1, \eta_2, \ldots, \eta_m)'$ inside a $(1-\alpha)$ confidence region, then the robustness measure for overall desirability is defined as:

$$\begin{aligned} D_R(x) &= \min_{\eta} D[d_1(\eta_1), d_2(\eta_2), \ldots, d_m(\eta_m)]|(\eta_1, \eta_2, \ldots, \eta_m) \\ &= \min_{\eta} \Big\{ \prod_{i=1}^{m} (d_i(\eta_i)^{w_i})^{\frac{1}{\sum w_i}} |(\eta_1, \eta_2, \ldots, \eta_m) \Big\}. \end{aligned}$$

This thus represents the worst-case scenario; the minimal overall desirability D value when the true responses are located anywhere in the confidence region. The goal here is to find the factor settings that are associated with the best worst-case scenario.

Let $min_{eta_i} d_i(\eta_i)|\eta_i \in [LC, UC]$ be denoted as d_{R_i}, called the individual robust desirability. Then, $d_{R_i}(x) = d_i[LC_i]$ for Larger the Better responses, $d_{R_i}(x) = d_i[UC_i]$ for Smaller the Better responses, and $d_{R_i}(x) = d_i[LC_i]$if $|LC_i - T_i| \geq |UC_i - T_i|$, or $d_i[UC_i]$ otherwise for Nominal the Better responses. $D_R(x)$ can then be simplified using the confidence region to:

$$D_R(x) = (\prod_{i=1}^{m} min_{\eta_i} d_{R_i}^{w_i})^{\frac{1}{\sum w_i}}.$$

He et al. suggested finding the robust optima associated with a gradually decreasing set of α values to aid in finding the best solution with some level of flexibility. The authors presented an example searching operating conditions that optimize simultaneously the yield, viscosity, and molecular weight of a chemical process. For comparison, the traditional Derringer's desirability function was utilized (i.e. setting $\alpha = 1$) and shown to have a weakness in that the output responses' confidence intervals are not always in a specified allowed range. He's method reveals a robust solution based on confidence intervals of output responses rather than individual predicted values.

3.3.6 Minimization of Variance of Predicted Responses using Desirability Functions (Chen et al., 2012)

While He et al. (2012) sought to maximize the minimal overall desirability D value when the true responses are located anywhere in a confidence region, Chen et al. (2012) sought to narrow prediction intervals by minimizing the variability in predicted responses, an objective not achieved by traditional desirability methods [17].

Let $var(\hat{y}_i)$ and $v_i(x)$ be defined as in Section 1.4.4. The variance of the predicted response, $var(\hat{y}_i)$, can then be minimized if the standard deviation (SD) of each predicted response is transformed into an individual desirability function d_{s_i}. The SD of predicted responses is certainly of the STB type, so the lower bound L is set to be 0. The upper bound can be $c_i\sigma_i$ (where c_i is a multiple selected by the user), where smaller values of c_i allow for smaller ranges for the values of $sd(\hat{y}_i)$.

Chen et al. call the desirability function for the $sd(\hat{y}_i)$ the secondary individual desirability function, defined as:

$$d_{s_i} = \begin{cases} (\frac{c_i\sigma_i - sd(\hat{y}_i)}{c_i\sigma_i})^r = (\frac{c_i - \sqrt{v_i(x)}}{c_i})^r & \text{for } 0 < v_i(x) < c_i^2 \\ 0 & \text{for } v_i(x) \geq c_i^2. \end{cases}$$

The authors point out that if x^* denotes the optimal solution from Harrington's desirability function, defining $c_i = \sqrt{v_i(x^*)}$ would lead to the secondary individual desirability function to represent the relative change in standard deviation between the new optimal solution and x^* when scaling parameter $r = 1$.

A combination of all the secondary individual desirabilities into a secondary overall desirability function can then be done by taking the geometric mean:

$$S = (d_{s_1} d_{s_2} \dots d_{s_m})^{1/m}.$$

The augmented desirability function is then defined as the weighted product of the overall desirability score D and the secondary overall desirability S:

$$DS_\lambda = D^\lambda S^{1-\lambda},$$

where λ and $1-\lambda$ are user-selected weights between 0 and 1 that indicate the relative importance of optimizing D versus S. Lower values of λ place less importance on the optimization of the multiple responses. Contour plots can be used to graphically display how optimal solutions shift with varying values of λ.

In a study to evaluate performance of the augmented approach in DS_λ, two applications were examined. In one, the optimal factor settings to maximize extraction yields of three different saikosaponins were sought. In a second, the ideal settings for voltage and buffer solution concentration were identified to optimize resolutions, analysis time, and capillary current. Utilization of the authors' augmented method resulted in a reduction of variances of predicted responses by 50%.

3.4 AIDS Case Study

In this subsection we apply desirability functions to assess HIV disease progression in a real study. Desirability functions are a useful tool in the assessment of HIV because looking at a single covariate such as CD4+ cell count is insufficient to understand the overall activity of the virus in the body. The goal is to simultaneously assess several covariates contributing to the course of the disease, including CD4+ cells, which are white blood cells that find and destroy bacteria and viruses in the body, and CD8+ T-cells, which recognize infection and kill virally infected cells.

The MACS Study began in 1984, enrolling 4,954 gay and bisexual men and following-up semi-annually [59]. Since then, the MACS Study has grown to include newer cohorts. We focus on the 2001-2003 "new recruit cohort", which enrolled 1,350 men between October 2001 and August 2003 [58]. We will keep the originally given visit names in this analysis; visit 365 was called the initial visit by the MACS clinicians for this cohort.

The responses are hematocrit (%), CD4+ cell count, CD8+ cell count, and white blood cell count in construction of overall desirability. Hematocrit is the ratio of the volume of red blood cells to the total volume of blood. Viral load was not included in the analysis due to a high percentage of missing information - nearly 40% of the cohort did not have viral load recorded throughout the study. Older cohorts in the MACS dataset had even less viral load data. One later cohort exists beginning in 2010, but has less subjects and thus is not inspected for this analysis. Had more viral load data been available or had missing data patterns been fully understood in this study, viral load would have been a good example of a gold standard (d_g) to derive shape parameters and weights as in Section 3.2.9 (perhaps partitioned by time elapsed since infection, since the relationship between viral load and time is not linear).

Table 3.2 provides some details for the variables of interest and the parameters used in their respective individual desirability functions. The four variables are NTB variables. The midpoint of an acceptable range was set as the target value, and r_1 and r_2 were set to be values that yielded a 0.8 desirability for the edges of the variable's acceptable range. This allowed NTB variables to have a set of values that would score well between 0.8 and 1, with values deviating from the acceptable range scoring less than 0.8. Weights were arbitrarily set for sake of demonstration, giving most importance to CD4+ and CD8+ cell count. In practice, it is recommended that experienced experts of the disease are entrusted with defining the weight parameters of the individual components.

	Type	L	T	U	r	r_1	r_2	w
Hematocrit (%)	NTB	29	44	55	-	0.5	0.22	0.2
CD4+ cell count	NTB	200	1000	2402	-	0.25	0.5	0.3
CD8 cell count	NTB	0	575	3007	-	0.2	1.1	0.3
White blood cell count	NTB	0	7750	48700	-	0.4	2.4	0.1

Table 3.2: Outcomes of interest in assessing subject HIV disease progression, their individual desirability function definitions, and weights.

Each subject's lab values in a single visit contributed to an overall desirability score D at that time point, where D was calculated using the parameters from Table 3.2 in Equation B.11. Figure 3.4a shows the average desirability for all patients who were HIV-negative or positive at the specified visit. The overall desirability score across visits for HIV-negative subjects, depicted by the black line, remains relatively steady. The gradual increase in overall desirability from HIV-positive patients is attributed to two reasons: first, newly converted subjects take a few to several visits to have a dramatic decline in scores; second, the efficacy of antiviral treatment pulls up the desirability score of subjects who have had HIV for longer periods. These two reasons are supported by Figure 3.4c, to be discussed shortly.

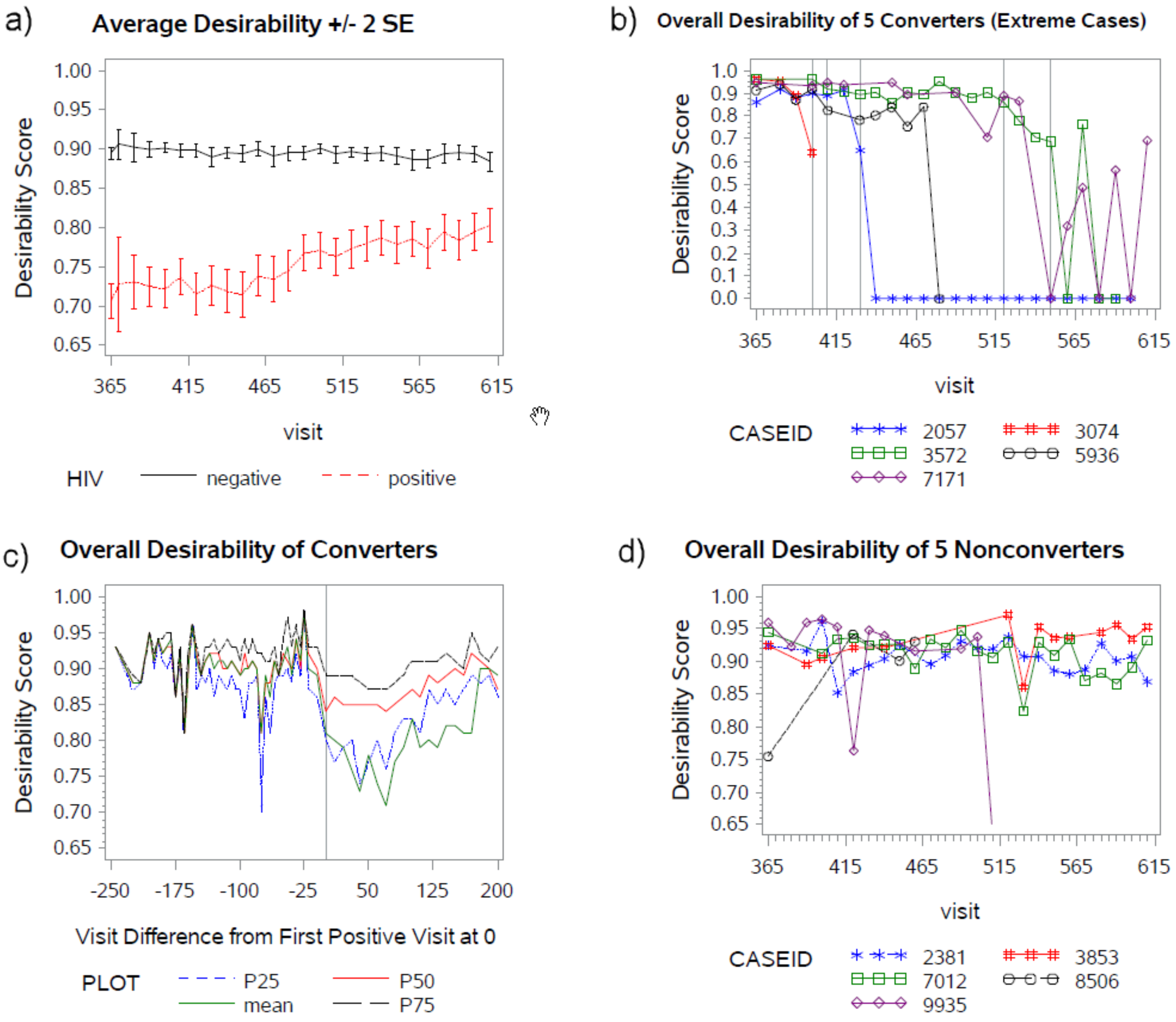


Figure 3.4: MACS study: a) average overall desirability score ± 2SE across visits for all HIV- negative and positive subjects; b) overall desirability score of five extreme cases of seroconverters; c) overall desirability score over time for seroconverters with the diagnosis visit calibrated as time 0; d) overall desirability score over time for nonconverters.

Figure 3.4b reveals desirability scores over time for five individuals who became HIV positive during the study. The five cases shown were selected due to their extreme changes in overall desirability score. Patient 3074 was first discovered to be HIV positive on visit 400, 5936 on visit 410, 2057 on visit 430, 3572 on visit 520, and 7171 on visit 550, as indicated by the gray vertical reference lines in the plot. The corresponding drops in their desirability scores reflect the ability of the desirability score to capture patient status. The figure demonstrates that a subject's overall desirability score tends to decline before the subject is officially diagnosed with HIV - see the purple line connecting the diamonds for subject 7171, whose score declined from nearly 0.9 at visit 490 before dropping to just over 0.7 at visit 500 and rising shortly back to nearly 0.9 before plummeting to 0 at the next visit of 550, where he was formally defined as a seroconverter.

Figure 3.4c shows overall desirability over time *relative to the first sero-positive visit*, which we reset as zero and indicate with a gray vertical reference line. This figure provides some clues to the overall positive trend of overall desirability in the subjects who are HIV-positive at a given visit in Figure 3.4a. The overall desirability score visibly declines at about the diagnosis visit minus 30. The desirability score continues to decline until about diagnosis visit plus 75, showing improvement in patients' lab values. Figure 3.4d shows overall desirability of five randomly selected individuals who remained HIV negative during the study. The plummeting desirability for patient 9935 was attributed to a sudden drop in hematocrit to 29%. In general, the nature of the overall desirability score in nonconverters is stable within the range of 0.85 and 0.95 relative to that of converters.

This case study demonstrates the value of utilizing desirability functions in assessing health outcomes. It also is an example of how desirability scores can still provide valuable insight in health outcomes even when the desirability functions are defined without a gold standard.

3.5 Advanced Programming & Algorithms

After an overall desirability function is defined, one popular problem to be solved is to identify the value x of each of p input variables that maximize the overall desirability function while still remaining within a specific plausible range $[L(x_h), U(x_h)]$:

$$\max D = (d_1^{w_1} d_2^{w_2} ... d_m^{w_m})^{\frac{1}{\sum_{i=1}^{m} w_i}}$$
$$\text{s.t. } L(x_h) \leq x_h \leq U(x_h), \qquad h = 1, 2, \dots, p.$$

Different optimization algorithms have been used to do so. For example, Derringer and Suich utilized a direct search method known as Hook-Jeeves [26]. The strength of the Hook-Jeeves search method is that it need not be differentiable; its weakness lies in its high probability of identifying only local optimal solutions. Castillo used a gradient-based algorithm, which required differentiable functions as discussed in Section 3.1.3, and also yields local optimal solutions sensitive to the initial search point [25]. Genetic algorithms are a search technique robust to initial search parameters that can find global maximums [44].

Genetic algorithms (GAs) are inspired by natural selection as observed in evolution and are gaining popularity in the area of optimization. GAs were first introduced by John H. Holland in 1960, and have been further expored by David E.Goldberg [44, 34]. "Living organisms are consummate problem solves. They exhibit a versatility that puts the best computer programs to shame," Holland prefaced in his courses [44]. GAs search stochastically through the real space of the problem by generating a random initial population.

This starting point of a set of random initial "population" - or, solutions - is what differentiates GAs from other conventional search and optimization techniques. Within the initial population, each individual's fitness as determined by chromosomes is assessed so that the fittest individuals (solutions) can reproduce the next generation. Chromosomes represent solutions to an optimization problem, whereby the solution's set of values are denoted by $x_1, \ldots, x_p$. Selection, crossover, and mutation are examples of genetic operators that will affect the reproduction cycle and generation of subsequent generations. New generations' members continue to be assessed and mated until an optimal solution is found, whereby the algorithm stops. A solution is deemed optimal when no improvement in fitness is seen across several consecutive generations. The change in fitness considered insignificant and the number of generations considered sufficient to declare a solution optimal varies depending on the problem and the preferences of the user.

To initiate the genetic algorithm, one must first specify:

- Population size: the number of chromosomes (scenarios), N, that are retained in each generation.
- Number of replications: the number of times, v, each scenario is simulated.
- Crossover rate: the probability of crossover, P_c.
- Mutation rate: the probability of mutation, P_m.

The total desirability value of scenario j depends on the response variables i in scenario j in replication v, y_{ijv}, with $i = 1, \ldots, p$, $j = 1, \ldots, N$, $v = 1, \ldots, n$, where the responses resulting from some input variable x_{ij} with $i = 1, \ldots, p$ and $j = 1, \ldots, N$. Individual desirability scores d_{ijv} are calculated for each scenario's responses in each replication, which thereby contribute to the overall desirability scores of scenario j in replication v: D_{jv}. The mean overall desirability of each scenario $\bar{D}_j$ is then calculated by averaging D_{jv} across the number of replications $v = 1, \ldots n$.

To employ the GA, one must understand the crossover and mutation genetic operators. The crossover process mates pairs of chromosomes by randomly selecting a pair of chromosomes A and B from the current generation with probability P_c. The chromosomes are built of genes $[a_1, a_2, \ldots, a_p]$ and $[b_1, b_2, \ldots, b_p]$, respectively. This pair of chromosomes A and B will reproduce the new chromosomes A_{new} and B_{new} via a parameter λ ranging between 0 and 1 and crossover equations:

$$A_{\text{new}} = \lambda A + (1 - \lambda) B, \tag{3.5}$$

$$B_{\text{new}} = (1 - \lambda) A + \lambda B. \tag{3.6}$$

Mutation, on the other hand, replaces a gene a_j with a new gene a_j* via the mutation Equations 3.7, 3.8.

Let l_j and u_j be lower and upper limits of the specified gene a_j, u^* be a uniform random variable between 0 and 1, i be the number of the current generation, and maxgen be the maximum number of generations.

$$a_j* = a_j + (u_j - a_j) \times u^* \times (1 - \frac{i}{\text{max gen}}), \tag{3.7}$$

$$a_j* = a_j + (a_j - l_j) \times u^* \times (1 - \frac{i}{\text{max gen}}). \tag{3.8}$$

Note that the term $1 - \frac{i}{\text{max gen}}$ has a value close to 1 in the first generation, and a value close to 0 in the last generation, resulting in larger mutations in early generations, and almost no mutation in the last generations.

How are a pair of chromosomes A and B selected for crossover and mutation? Pasandideh (2006) discusses and compares four methods relevant to desirability functions, finding them not statistically different [66]. The four methods all utilize the mean desirability $\bar{D}_j$ of scenario j. Since the results yielded from the four methods are similar, we will review just one method for chromosome selection. In this method, a multiple-comparison statistical test is used to adjust for the random nature of the desirability, grouping chromosomes such that there is no statistical difference within the groups, but there exist differences among different groups. The critical value of Least Significant Difference (LSD) is calculated as $LSD = t_{\alpha/2,N(n-1)} \times \frac{\sqrt{2MSE(\bar{D}_j)}}{\sqrt{n}}$. Any pairs of $\bar{D}_j$ that differ less than the LSD are considered statistically equal. The chromosomes $x_1, \ldots, x_p$ are arranged in ascending order, grouping them together when they differ by less than the LSD: when they are considered statistically equal. Each group k of chromosomes is then selected with probability

$$p_k = \frac{\sum_{\forall j \in \text{Group k}} \bar{D}_j}{\sum_{j=1}^{N} \bar{D}_j},$$

whereby if a group is selected, its best chromosome as defined as having the highest $\bar{D}_j$ is selected for mating. N selections of chromosomes are made to make a generation with N chromosomes. The crossover and mutation operators of Equations 3.5 - 3.8 are applied to the N chromosomes. The algorithm is repeated until the stopping criteria is reached.

The reader is encouraged to refer to the original work for further details. Examples of utilizing the genetic algorithm to optimize the inputs or components of an overall desirability score can be found in Kim (2004), who applied the method to optimization of a gas metal arc welding process [52], and Pasandideh (2006), who highlighted the value of the method with simulated numerical results [66].

3.6 Discussion

In this chapter, we have provided an overview on desirability functions and how they are widely used in various disciplines to assess the overall quality of a product or outcome by combining several outcomes of interest into an overall composite score. The flexibility of desirability functions makes them an appealing tool for assessing health outcomes. Interestingly, desirability scores have been used only in a handful of medical applications and applied in a simple way. For example, the default shape parameters of one and equal weighting are generally assumed in medical applications even though these parameters can be estimated using nonlinear least squares and the Lp norm to assess an overall outcome over time. Medical applications also have yet to regularly incorporate model uncertainty into the desirability scores, or use functions to minimize variance of predicted responses. Despite the simple applications of desirability scoring used in medical applications thus far, the examples presented in this chapter highlight the ease of simultaneous interpretation of multiple responses reflected in the desirability scores.

The next chapter expands upon the work of Schindler discussed in Section 3.2.8 by employing desirability functions to assess a wider variety of design components and design types. The value of utilizing desirability functions as a tool in clinical trial design quality assessment will be demonstrated through a handful of case studies.

Chapter 4

Evaluation of Clinical Trial Design Quality Using Desirability Functions

The aim of Chapter 2 was to minimize overall total expected responses in two-arm trials with correlated responses between the two treatment groups. It could be seen that achieving this objective might come at the cost of other objectives. This chapter extends upon the ideas of Schindler from Section 3.2.8 by creating a framework to simultaneously evaluate multiple characteristics of a design. Section 4.1 discusses examples of the different components or characteristics of a clinical trial that contribute to its quality. With the exception of Type I error and power, the components included in this chapter are unique and novel components of desirability functions in the assessment of trial quality. Section 4.2 discusses how scale parameters and weights can be selected during the construction of desirability functions used to evaluate clinical trial design quality. A proposed framework is then presented in Section 4.3 regarding the construction of individual desirability scores for the characteristics under review during assessment of candidate clinical trial designs. The chapter ends with Sections 4.3.1 through 4.3.3, which provide examples of the application of the proposed framework. These examples include evaluation of two-arm clinical trials with binomial responses, continuous responses, and correlated continuous responses between the treatment arms. The value of utilizing desirability functions to pinpoint the strengths and weaknesses of a design and to culminate this information into a well-informed decision is shown. We have created an online tool to help trialists implement the framework presented in this chapter at `https://priscillakyen.shinyapps.io/DesignEvaluation_beta/`.

4.1 Components of Designs Considered

In the beginning phases of designing a clinical trial, research, ethical, and logistical objectives are formed. Most clinical trials aim to fulfill multiple objectives, and some of these objectives may compete [55, 41, 88]. Examples of competing objectives include: a) sufficient sample size and power vs. cost; and b) sufficient treatment group sample sizes to make convincing, unbiased treatment group comparisons vs. minimizing allocation of patients to a potentially inferior or toxic treatment. Competing objectives are indeed a large challenge during clinical trial design selection. We proceed to review design components to be evaluated in the context of this chapter.

4.1.1 Treatment Group Size Imbalance

Forced balance procedures such as those discussed in Section 1.1.1 are used for balancing treatment assignments. Balanced group sizes in clinical trials hedges against accidental bias, the bias in the estimated treatment effect that is due to the omission of significant confounders from the model [55]. This shall be shown in Section 4.1.2. Friedman et al. (1981) argue that equal allocation amongst treatment groups has two benefits: first, power is maximized when allocation is equal; second, equal allocation caters to the ethical concept of equipoise that clinical trialists should believe to be true at the start of a trial [32]. The potential imbalance of non forced balance designs such as CRD and adaptive designs may reduce power. Lachin has shown, however, that allocation ratios must exceed 7:3 in order to severely impact power [54]. Recent literature has convinced many clinical trialists that the first benefit of maximum power mentioned by Friedman et al. (1981) is not necessarily true ([73, 46]). However, the culture of clinical trials has embedded the concept of treatment group size balance as an important part of clinical trial design.

In this chapter, each design is evaluated using 10,000 simulated trials. The treatment group size imbalance, defined as $n_E - n_C$, is assessed at the end of each simulated trial. The distribution of observed treatment group size imbalance for each given design is then assessed using an individual desirability function.

4.1.2 Accidental Bias

Accidental bias describes the measure of bias in the treatment effect that is introduced due to some unobserved yet confounding covariate.

In this subsection, we shall modify notation and allow $T_j = 1$ for an individual in experimental arm E, and $T_j =$ -1 for an individual in control arm C, $i = 1, ..., n$.

Let us consider the true model to be a standard normal error regression model:

$$E(Y) = \mu e + \alpha t + \beta z, \tag{4.1}$$

where e is a vector of ones: $e = (1, 1, ..., 1)'$, t is the treatment vector given by $t = T = (t_1, ..., t_n)'$, and z is a covariate that is significantly associated with the outcome Y.

Denoting the design matrix X, we see that

$$X = \begin{bmatrix} 1 & t_1 & z_1 \\ 1 & t_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & z_n \end{bmatrix}, \qquad X'Y = \begin{bmatrix} e'Y \\ t'Y \\ z'Y \end{bmatrix}$$

.

Using ordinary least squares method, if we look at $(X'X)^{-1}X'Y$, then the consistent estimate of α is

$$E(\hat{\alpha}) = \frac{n(\mu e't + n\alpha + \beta z't) - e't(n\mu + \alpha e't)}{n^{-}(e't)^2}.$$

However, if the covariate z is incorrectly excluded from the model of Equation 4.1, then

$$E(Y) = \mu e + \alpha t,$$

Denoting the design matrix X, we see that

$$X = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}, \qquad X'Y = \begin{bmatrix} e'Y \\ t'Y \end{bmatrix}$$

.

$$(X'X)^{-1} = \frac{1}{n^2 - (e't)^2} \begin{bmatrix} n & -e't \\ -e't & n \end{bmatrix}$$

.

Using ordinary least squares method, if we look at $(X'X)^{-1}X'Y$, then the biased estimate is

$$\hat{\alpha} = \frac{nt'Y - (e't)(e'Y)}{n^2 - (e't)^2}.$$

The squared bias term is then

$$[E(\hat{\alpha} - \alpha)]^2 = (\frac{n}{n^2 - (e't)^2})^2 \beta^2 (z't)^2.$$

The impact of imbalanced treatment group sizes is highlighted by the $(e't)$ and is clear: larger imbalances contribute to greater bias in the estimate of the treatment effect. The bias in the estimate of the treatment effect also increases with the magnitude of the coefficient β for the omitted covariate z. Lastly, accidental bias depends on the term $(z't)^2$, which is zero when z is orthogonal to t. The unconditional expectation can be taken for a fixed vector z, with t being a realization of T and $\Sigma_T = Var(T)$:

$$E(z'T)^2 = z'\Sigma_T z,$$

By Rao, $E(z'T)^2$ cannot exceed the maximum eigenvalue of Σ_T ([71], p62). Due to this inequality, Efron uses the maximum eigenvalue of Σ_T as a criterion to evaluate the degree to which accidental bias impacts a design. Specifically, the maximum eigenvalue of Σ_T can be used as a minimax criterion, such that the randomization procedure $T_1, \ldots, T_n$ with the minimum maximum eigenvalue is selected [28].

In the context of this work, the asymptotic derivations for the covariance vector of T are not ideal. Instead, my approach is to calculate an accidental bias factor:

$$\text{Accidental Bias Factor Estimate } = \left(\frac{n}{n^2 - (e't)^2}\right)^2 \hat{\lambda_{max}}, \tag{4.2}$$

where λ_{max} is the maximum value of the covariance matrix of T:

$$\text{Var } T = \mathbf{\Sigma} = \begin{bmatrix} E[(T_1 - ET_1)(T_1 - ET_1)] & E[(T_1 - ET_1)(T_2 - ET_2)] & \ldots & E[(T_1 - ET_1)(T_n - ET_n)] \\ E[(T_2 - ET_2)(T_1 - ET_1)] & E[(T_2 - ET_2)(T_2 - ET_2)] & \ldots & E[(T_2 - ET_2)(T_n - ET_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(T_n - ET_n)(T_1 - ET_1)] & E[(T_n - ET_n)(T_2 - ET_2)] & \ldots & E[(T_n - ET_n)(T_n - ET_n)] \end{bmatrix}$$

The expected value of a treatment for patient j is estimated in my simulations by taking the mean of the patient's treatment indicator value across all iterations. Specifically, for patient j and iteration i $(i = 1, ..., \text{iter})$, $ET_i = \frac{\sum_{i=1}^{\text{iter}} T_i}{\text{iter}}$. This estimate then allows us to find the estimate $\hat{\mathbf{\Sigma}}$, which is then used to find an estimate of the accidental bias factor. To my knowledge, previous simulations discussed in literature look only at λ_{max}, whereas this work newly incorporates treatment group size imbalance - which also contributes to accidental bias - through simulation.

As the number of iterations $iter \to \infty$, the estimate of the accidental bias factor reaches its theoretical value. Designs yielding lower accidental bias factor estimates are favorable. In practice, this means the designs with lower average treatment group imbalance, and with less variability within treatment assignments for each subject j, are less likely to have treatment effect estimates impacted by unobserved covariates.

Lachin et al. note that with the exception of truncated binomial design, accidental bias seems to be negligible for forced balanced designs [55]. However, accidental bias in response-adaptive designs seems to be an area less studied. Using the accidental bias factor estimate from Equation B.26, this work is able to compare the impact of unobserved covariates on a broader range of designs.

4.1.3 Covariate Imbalance

Covariate imbalance occurs when the average values of a potential confounder C are very different between treatment groups. Covariate imbalance is not used to describe differences in response Y. Imbalance in covariates is a non-ideal scenario that statisticians try to avoid during the design stage. Instead, covariate balance is desired because balance across all potential confounders C will assure analysts that any differences between the two treatment groups at the end of a trial can be attributed to the treatment effect, and not to some other lurking variable. Most clinical trials report a large number of baseline covariates by treatment group to ensure that the two groups were "balanced" at baseline, and that differences at the end of the trial should thereby be attributed to the differences in treatment.

In this work, the setup for the study of covariate imbalance follows that of Lachin and Rosenberger (2016) [55]. Three different patterns are studied to see the probability of imbalance at the end of a trial:

1. $C1_1, C1_2, ..., C1_n$ are i.i.d. N(0,1).
2. $C2_1, C2_2, ..., C2_n$ drift linearly over time on the interval (-2,2] + a N(0,1) random variable.
3. $C3_1, C3_2, ..., C3_n$ are autocorrelated. Specifically, $C3_j = C3_{j-1}$ + a N(0,1) random variable, with j = 2,...,n and the first of the series $C3_1$ equaling a N(0,1) random variable.

These three scenarios represent three different types of covariates. The first is standardized normal, and is a good representation of what one expects from most covariates measured in a study. Different means and variances can be simulated, but N(0,1) is a representative proxy. The second scenario is representative of a covariate subject to a linear time trend. This is not to be confused with a linear time trend influencing the primary outcome of interest Y of a trial. An example of a covariate subject to a linear drift may be improving average blood pressure in patient population, when blood pressure is not the primary endpoint of interest. The third scenario is for autocorrelated variables, also known as *serially correlated* or *serially dependent*. This means that a covariate value is not independently and identically distributed, but rather, depends on the previous value. Returns on stock prices are frequently used as an example of autocorrelated variables, since past returns seem to influence future performance and returns. Other examples include annual rainfall, sunspot activity, and the price of agricultural products. In health, autocorrelation is seen in covariates quantifying exposure to pollutants. For example, asthma symptoms and daily ambient particulate matter concentrations are characterized as being related through an autocorrelated lag model [74].

At the end of 10,000 simulated trials for a design, we compute the frequency in which $|\overline{C}_A - \overline{C}_B| > \epsilon$, where ϵ can be specified by the user. The (frequency/total number of iterations) yields the simulated probability of covariate imbalance. Although not explored in this work, a trialist may simulate different covariates other than C1, C2, C3, and may vary ϵ depending on the threshold of imbalance acceptance for a particular covariate.

4.1.4 Selection Bias

Selection bias stems from the ability of an investigator or experimenter to predict the next treatment assignment, and therefore possibly bias the results by removing the random allocation of patients and the independence of patient characteristics and treatment assignment [15]. If the investigator has information on prior assignments, s/he may introduce bias by allocating a patient when the next assignment is likely to be the desired assignment from the perspective of the investigator. The bias may also take form when deciding against a prescribed allocation.

Since Blackwell & Hodges' introduction of this topic in 1957, Berger (2005) has further categorized selection bias into four types [15, 9]. First-order selection bias occurs when either the patient or the investigator is able to choose a specific treatment group, knowing the patient's characteristics. This is common in design's using Zelen's method, as discussed in Section 2.5. Second-order selection bias occurs when a clinical trial's investigator has access to the trial's randomization list and has the potential to enroll or exclude patients in a specific order. Third-order selection bias occurs when only future patient allocations are concealed, thus giving the investigator the ability to predict future allocations based on prior assignments. Fourth-order selection bias, also known as residual selection bias, arises when investigators are blinded to both past and future allocations.

The work in this dissertation focuses on third-order selection bias. In practice, a clinical trial usually targets double-blinded treatment assignment (the individual responsible for assigning treatments and the analysts do not know who receives which treatment). However, in reality, selection bias is still a risk that should be considered during the design of a trial. In spite of the double-blinded nature of many trials, treatment assignment may sometimes still be guessed or even obvious: patients in different treatment arms may experience different side effects; or sometimes the treatment arms themselves are unable to be masked (e.g. surgery vs. chemotherapy). In a meta-analysis of randomized clinical trials in which surgery was an intervention, less than 25 % of the trials concealed treatment allocation [49]. Even in nonsurgical cases, Berger stated his opinion that target allocations in RAR designs that minimize a specific objective function come with the price of increased selection bias. In a response to Taves's statement in 2010 claiming that "Minimization should be the method of choice in assigning subjects in all clinical trials" [81], Berger wrote:

"The idea behind minimization is brilliant, but its failure is the flip side of the same coin, and cannot be separated from its benefit. The claim to fame is that subjects are allocated not randomly but rather deterministically, so as to minimize an imbalance function (hence the name). It is this, the heart and soul of the method (and not some tangential aspect), that leads to its downfall.

"Investigators using minimization can determine the group to which a prospective subject would be allocated, and then decide if this is a good thing or a bad thing, in terms of creating an imbalance with respect to some key predictor or outcome not considered in the imbalance function. In other words, there is no allocation concealment, and selection bias is there for the taking. [Even by removing a patient's necessary allocation to a treatment, but rather assigning a high probability], are we really to believe that a betting man needs certainty, and not just good odds, to place a wager? If the odds of each treatment come close enough to 50% so as to truly take a bite out of selection bias, then the baby is lost with the bath water, and minimization no longer does what it purports to do. There is simply no way around this."

- Berger, in a letter to the Editor of Contemporary Clinical Trials [10].

The literature has shown that systematic baseline imbalances can occur in randomized trials, where selection bias would force imbalance in covariates influencing patient allocation [76, 82]. Berger (2005) states that randomization is necessary to ensure that any observed baseline imbalances are random, but it is not sufficient [9].

The predictability of a randomization sequence is given by

$$\rho_{pred} = \sum_{j=1}^{n} E\left|[E(T_j|F_{(j-1)})] - \frac{1}{2}\right|,$$

which is a measure in restricted randomization of the difference between the conditional probability of treatment assignment and the conditional probability [8]. We can then calculate a distribution of observed selection ρ_{pred} values across all simulations.

4.1.5 Chronological Bias

Chronological bias is contained within the larger umbrella of accidental bias and is an area of focus both in the literature and in this work. This is because the fulfillment of objectives such as minimizing treatment failures or minimizing the number of patients allocated to an inferior treatment is achieved by targeting one of the target allocations discussed in Section B, which can result in a large number of consecutive patients being enrolled to a single arm. This can be a problem when the performance of a patient is associated with time. It is worth differentiating here that chronological bias speaks to the time trends mentioned in this chapter, which are different than covariate C2 of section 4.1.3, since the former trend affects the response or endpoint of interest, and the latter affects an independent variable or confounder.

Three time trends studied in this work are linear, logarithmic, and stepwise in nature. A linear time trend signifies a linear change in the expected outcome of a patient across the course of the trial. For example, studies have noted that there is "good evidence of more aggressive treatment of blood pressure in recent trials. Mean baseline systolic blood pressure in [the] ACTIVE [trial] was 6-8 mm Hg lower than in SPAF3, despite similar enrollment criteria" [38]. A second example of a linear time trend is in the multicentre randomized trial comparing azathioprine versus placebo for patients with primary biliary cirrhosis (PBC), where it was noted that there was a steady decline in baseline log serum bilirubin levels with time, thus leading to the conclusion that subjects enrolled later in the trial were on average healthier than those enrolled at the beginning [3]. An example of a logarithmic time trend is in surgical trials, when surgeons' operational skills improve over the course of the trial [36]. On the other hand, a stepwise time trend signifies a sudden change during the trial which changes the expected outcome of patients from that point forward. Examples of this in practice might be a single change in medical or administrative staff, an adjustment in enrollment criteria, or a natural disaster partway through the trial which changes the expected response of a subject.

To protect a randomized clinical trial against chronological bias, an analyst must consider the observed bias in the analysis. However, the bias is often unknown during the design of the trial. At that early stage, the best protection is choosing a randomization scheme that hedges against large potential chronological biases and their effects on treatment effect estimation. For example, the ICH E9 (1998) guidelines recommend randomization in blocks (e.g. Random Block Design) when chronological bias is a concern [37]. Blocks protect against time trends since they avoid large runs of enrolled patients being allocated to a single arm, and blocks of subjects are influenced by a time trend equally.

In the literature, Schindler (2016) and Tamm and Hilgers (2014) have studied linear, logarithmic, and stepwise trends over the course of the trial. Specifically, their setup is an extension of Rosenkranz (2011). The basic regression model that is the foundation for estimation of the treatment effect β_1 throughout this dissertation is shown in the box below. Let

Basic Regression Model Components

$$\boldsymbol{X} = \begin{bmatrix} 1 & T_1 \\ 1 & T_2 \\ \vdots & \vdots \\ 1 & T_j \\ \vdots & \vdots \\ 1 & T_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \tag{4.3}$$

where $T_j = \mathbb{1}($ subject j is in experimental arm E$)$. Then, the response Y is modeled by:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\beta_{TIME} + \boldsymbol{\epsilon}. \tag{4.4}$$

The errors ϵ_j are independently and identically normally distributed with mean 0 and variance σ^2, and are assumed to be independent of treatment assignment. In this model, $\boldsymbol{Z}\beta_{TIME}$ reflects a time trend due to the sequential enrollment of patients which contributes to the expected response. Three time trends are studied: linear, logarithmic, and stepwise. The type of time trend influences the definition of the time covariate Z. The authors assume uniform enrollment throughout the recruitment stage of the trial, and thus use patient index number to model time throughout the trial. Assuming uniform enrollment and a time trend that influences subjects regardless of treatment group, $\boldsymbol{Z}$ is defined as:

1. Linear: $Z_j = (j-1)$
2. Logarithmic: $Z_j = \log(j)$
3. Stepwise: $Z_j = \mathbb{1}(j \geq t), 1 \leq t \leq n,$ (4.5)

where j indexes patient number in the order of enrollment, and β represents the strength of the time trend. For example, in a study with a linear time trend with $\beta = 1/127$ and $n = 128$, the expected response of subject j would be $E(Y_j) = \beta_0 + \beta_1 T_j + \frac{1}{127}(j-1)$. This means that the last enrolled patient ($j = 128$) would have an expected response 1 unit higher than the first enrolled patient ($j = 1$) simply due to the time trend, or "chronological bias". Similarly, subjects 1 through 127 would have linearly increasing expected outcomes by 1/127 units each, with the increase having no association with treatment effect.

The authors focus on evaluating how different permuted block randomization sequences perform with respect to chronological bias. The focus of interest is the bias and the variance of the estimate of the treatment effect. The authors also calculate the empirical type I error by simulating scenarios under the null hypothesis of no treatment effect, and evaluating the proportion of p-values that are less than $\alpha = 0.05$ when implementing a two-tailed t-test with pooled variance estimation.

Using their results, the authors recommend using small block sizes to hedge against biased estimates of the treatment effect. Lachin and Rosenberger (2016) comment that the impact of time trends in biasing the treatment effect estimate depends on the randomization procedure of a design, and consequently a worthy topic of exploration is which procedures mitigate the impact of chronological bias. Villar et al. (2017) look at the effects of time trends due to changes in standard of care and due to patient drift for RAR designs implemented in trials with binary responses [87]. Ryeznik et al. (2018) assess via simulation the performance of eight different designs, including CRD, PBD, DBCD, and urn designs, with three different doses, in the presence of time trends.

4.1.6 Expanding on Formerly Investigated Chronological Bias Patterns

This dissertation continues to expand from Tamm and Hilgers' work. The following changes which allow improved flexibility - and arguably accuracy - are implemented:

1. The time trends studied are not assumed to necessarily affect both treatment groups. A time trend may impact one group only, or the two treatment groups could witness different time trends or trends with different degrees of severity.

2. Two Type I errors are calculated, by simulating scenarios under the null hypothesis of no treatment effect, and computing the proportion of p-values that are less than $\alpha = 0.05$ in the Wald Test of a linear regression that a) does not include, and b) includes a covariate for time trend. Notice that a) is equivalent to the two-tailed t-test performed by Tamm and Hilgers. For more on Type I error, see Section 4.1.10.

3. Power is also evaluated, by simulating scenarios under the alternative hypothesis of an existing treatment effect, and computing the proportion of p-values that are less than $\alpha = 0.05$ in the Wald test of a linear regression that a) does not include, and b) includes a covariate for time trend. This approach is significant because including a covariate for time trend in the linear regression results in a loss in power. Statisticians may be interested in simulating the extra sample size needed to compensate for this loss. For more on power, see Section 4.1.10.

4. The performances of various designs not limited to Permuted Block Randomization are summarized.

5. A different patient enrollment pattern is studied: previous work has modeled time trends as a function of patient number; we now allow model patient recruitment with varying Poisson rates.

Regarding the first change, since a surgeon's improvement or lengthy trials are referred to as examples of sources of chronological bias, it may not be practical to assume that both treatment groups are subject to chronological bias, and if they were, subject to the same chronological bias. This thesis expands upon the previous literature by further exploring some scenarios that may be more likely in practical clinical trials.

We will use $\boldsymbol{Z}$ here to represent time trend covariates. The $\boldsymbol{Z}$ matrix of Equation 4.4 is adjusted to:

$$\boldsymbol{Z} = \begin{bmatrix} Z_{1\text{linEC}} & Z_{1\text{linE}} & Z_{1\text{linC}} & Z_{1\text{stepEC}} & Z_{1\text{stepE}} & Z_{1\text{stepC}} & Z_{1\text{logEC}} & Z_{1\text{logE}} & Z_{1\text{logC}} \\ Z_{2\text{linEC}} & Z_{2\text{linE}} & Z_{2\text{linC}} & Z_{2\text{stepEC}} & Z_{2\text{stepE}} & Z_{2\text{stepC}} & Z_{2\text{logEC}} & Z_{2\text{logE}} & Z_{2\text{logC}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{j\text{linEC}} & Z_{j\text{linE}} & Z_{j\text{linC}} & Z_{j\text{stepEC}} & Z_{j\text{stepE}} & Z_{j\text{stepC}} & Z_{j\text{logEC}} & Z_{j\text{logE}} & Z_{j\text{logC}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{n\text{linEC}} & Z_{n\text{linE}} & Z_{n\text{linC}} & Z_{n\text{stepEC}} & Z_{n\text{stepE}} & Z_{n\text{stepC}} & Z_{n\text{logEC}} & Z_{n\text{logE}} & Z_{n\text{logC}} \end{bmatrix}. \tag{4.6}$$

Let $\boldsymbol{\beta}_{TIME} =$

$$\boldsymbol{\beta_{TIME}} = (\beta_{\text{linEC}}, \beta_{\text{linE}}, \beta_{\text{linC}}, \beta_{\text{stepEC}}, \beta_{\text{stepE}}, \beta_{\text{stepC}}, \beta_{\text{logEC}}, \beta_{\text{logE}}, \beta_{\text{logC}})^T. \tag{4.7}$$

Then, responses Y are modeled as:

Response Model With Time Trend

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\beta_{TIME}} + \boldsymbol{\epsilon}. \tag{4.8}$$

If enrollment into the trial is uniform, the Z covariate is modeled as a function of patient number:

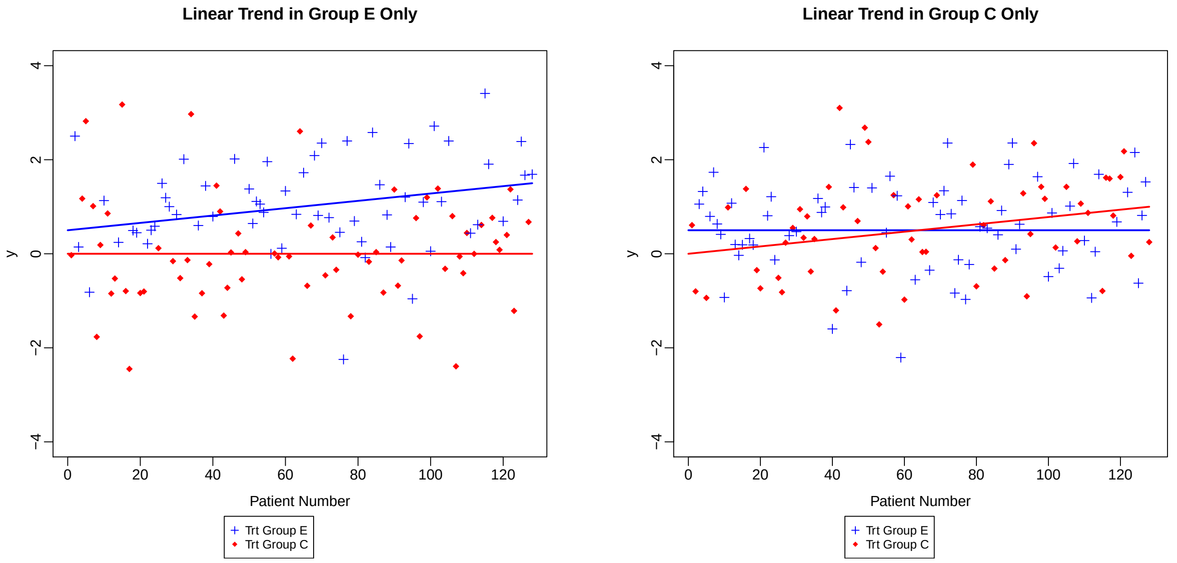
Definition of Z when subject enrollment is uniform

$Z_{j\text{linEC}} = (j-1)$, $Z_{j\text{linE}} = (j_E - 1)$, $Z_{j\text{linC}} = (j_C - 1)$,

$Z_{j\text{logEC}} = \log(j)$, $Z_{j\text{logE}} = \log(j_E)$, $Z_{j\text{logC}} = \log(j_C)$, and

$Z_{j\text{stepEC}} = \mathbb{1}(j > t)$, $Z_{j\text{stepE}} = \mathbb{1}(j_E > t)$, $Z_{j\text{stepC}} = \mathbb{1}(j_C > t)$.

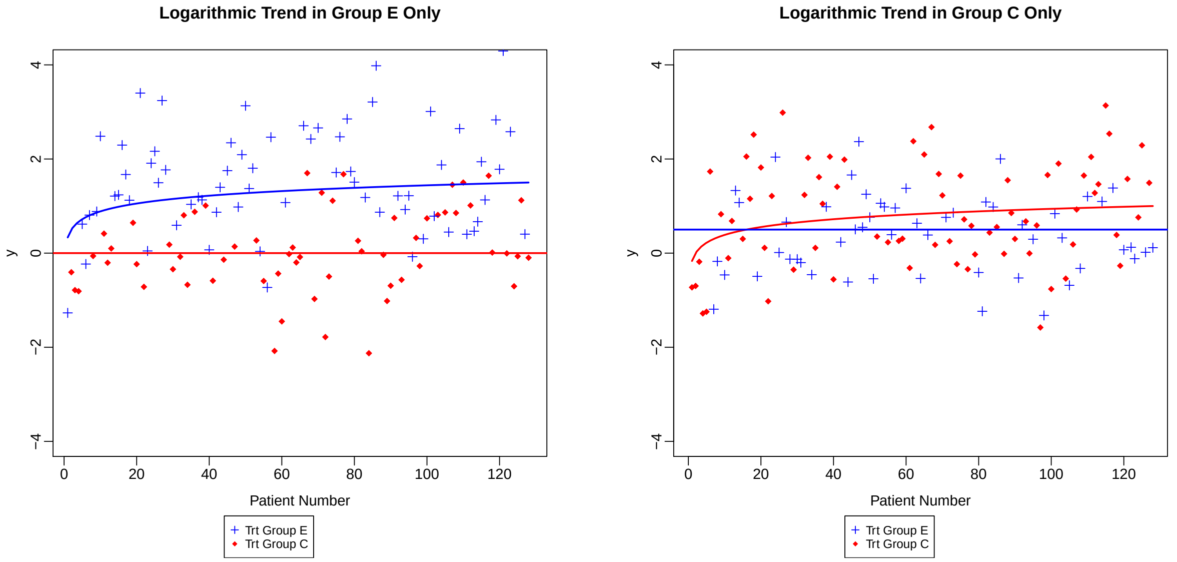
See Figures 4.1 through 4.3 for linear, logarithmic, and stepwise trends in a single treatment group only. The mean response in control group C is 0, and the mean response in experimental group E is 0.5. The strength of the time trend λ is chosen so that the last subject affected by the trend has an expected response



(a) $Z_{j\text{linE}} = (j_E - 1)$, $\beta_{\text{linE}} = 1/63$.

(b) $Z_{j\text{linC}} = (j_C - 1)$, $\beta_{\text{linC}} = 1/63$.

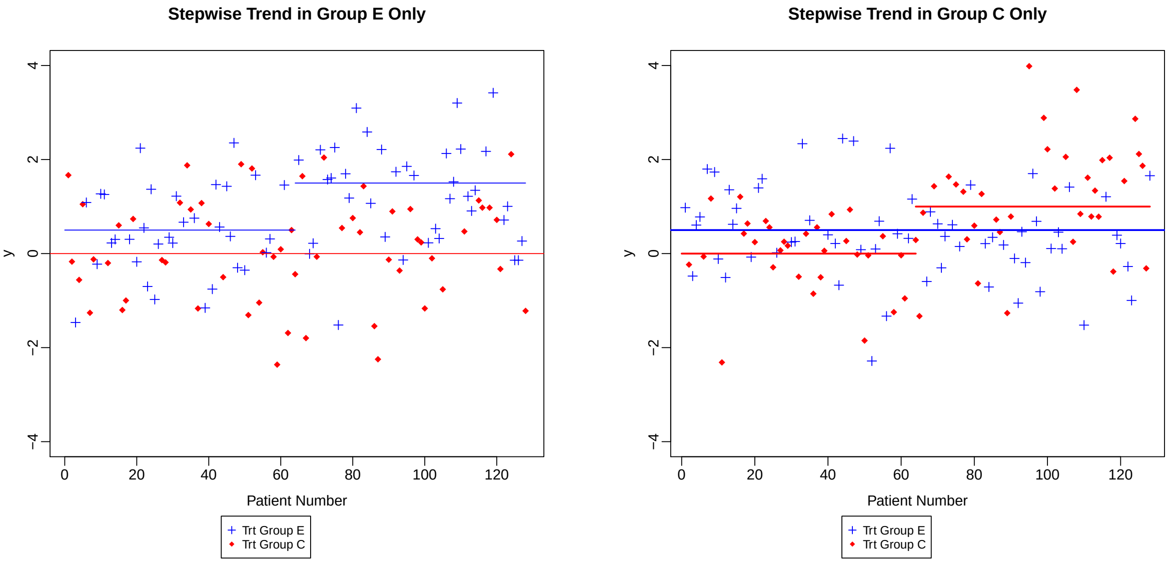
Figure 4.1: Linear trends across $n = 128$ subjects with 64 subjects in each arm, resulting in a net one unit increase in response from the first to last subjects affected by the time trend.



(a) $Z_{j\text{logE}} = \log(j_E)$, $\beta_{\text{logE}} = 1/\log(64)$.

(b) $Z_{j\text{logC}} = \log(j_C)$, $\beta_{\text{logC}} = 1/\log(64)$.

Figure 4.2: Logarithmic trends across $n = 128$ subjects with 64 subjects in each arm, resulting in a net one unit increase in response from the first to last subjects affected by the time trend.



(a) $Z_{j\text{stepE}} = \mathbb{1}(j_E \geq t)$, $\beta_{\text{stepE}} = 1$.

(b) $Z_{j\text{stepC}} = \mathbb{1}(j_C \geq t)$, $\beta_{\text{stepC}} = 1$.

Figure 4.3: Stepwise trends across $n = 128$ subjects with 64 subjects in each arm, $\lambda = 1$, $t = 33$, resulting in a one unit increase in response beginning at the 33rd subject of the arm affected by the time trend.

one unit higher than the first subject unaffected by the trend. The subjects enrolled between the first and last subjects are affected by the time trend in a linear, logarithmic, or stepwise function. For example, in Figure 4.3a, the mean response in the experimental group begins at 0.5, a indicated by the solid blue line, and the gap in response between the experimental and control group (solid red line) scan be attributed solely to the effect of the treatment. Halfway through the trial, a stepwise trend occurs, raising the expected response in the experimental arm from 0.5 to 1.5, resulting in the gap in response between the two groups to be attributed to both a treatment effect *and* an external stepwise trend. In Figure 4.1a and 4.2a, the start and end points are similar, but the pattern of the time trend achieves the final response in a linear and logarithmic fashion, respectively. When the trend is in the same direction as the treatment effect (e.g. Figures 4.1a, 4.2a, 4.3a), the treatment effect estimate will be overestimated, and Type I error will increase, since the response level can incorrectly be attributed to a response to treatment. When the trend is in the opposite direction as the treatment effect (e.g. Figures 4.1b, 4.2b, 4.3b), the treatment effect will be masked and harder to detect, resulting in a drop in power. For example, in Figure 4.3b, the average response in the control group (red lines) is equivalent to the average response in the experimental group (blue line), and no treatment effect would be detected if a Z covariate for stepwise trend were not adjusted for. The alternative hypothesis would be falsely rejected in this case.

In the definitions above, it is assumed that subjects are enrolled into the trial at a uniform rate. On the other hand, if enrollment into the trial is not uniform, measurement time m_j of the response plays a role in the definition of Z. Let $G \in [EC, E, C]$ represent whether the time trend affects both the experimental and control groups, the experimental group only, or the control group only. Let M equal the last measurement time, and $step \in [0, 1]$ indicate at what fraction of the total measurement time a stepwise change occurs.

Definition of Z when subject enrollment varies over time

$Z_{j\text{linG}} = m_j$,

$Z_{j\text{logG}} = \log(m_j)$, and $\qquad G \in [EC, E, C]$

$Z_{j\text{stepG}} = \mathbb{1}(m_j \geq step \times M)$.

Modling time trends a sa function of time rather than subject number provides flexibility and enhanced accuracy in our simulations.

When it is difficult to look at raw response values, Altman et al. suggest using CUSUM plots, which plots the cumulative sum S_j of response value less their expected value:

$$S_j = \sum_{j=1}^{n} (Y_j - Y_0), \tag{4.9}$$

where Y_0 is the expected value of the response [3]. There are multiple ways to define Y_0, including taking the average response across the first n_{start} observations, or letting it equal the specified expected response in the alternative hypothesis. In Figure 4.4, Y_0 is set to the expected value of the response as specified in the alternative hypothesis: 0 for responses in control group C, and 0.5 for responses in experimental arm E. If no time trends are present in the data, and the true mean is Y_0, we would expect the values of S_j to be close to zero, and the CUSUM plot should be nearly flat. A change in level of raw responses appears as a change in the slope of the CUSUM.

In Figure 4.4, a CUSUM graph replicates the trend reflected in 4.1a. It can be seen that the responses in the control arm C lie close to their expected value of 0, with a relatively flat line. On the other hand, the cumulative sum S_j in experimental arm E deviates further and further away from 0, with a positive slope. The changes in S_j from one patient in experimental arm E to the next point to the linear nature of the trend rather than a logarithmic or stepwise nature.

Simulating Recruitment

Two practical scenarios lead us to believe that chronological time trends are more accurately depicted when they are modeled as a function of time, rather than as a function of patient number. First, we would expect chronological trends to be a bigger concern when enrollment is slow. If there is buildup of resistance to an HIV vaccine over time, it would be less of a concern for a clinical trial that finished recruiting (and

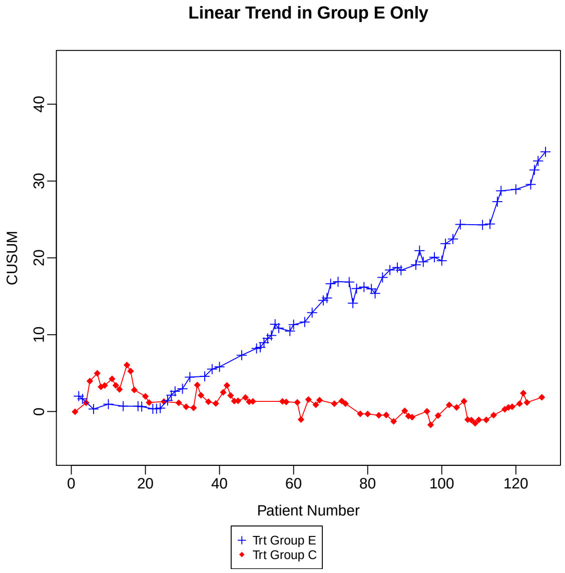


Figure 4.4: CUSUM plot displays the cumulative sum of response values less their expected value (see Equation 4.9) and depicts the same linear time trend in the experimental arm E as shown in Figure 4.1a.

measuring the primary endpoint of interest) in three months relative to one that took two years, even if both trials had the same sample size. Second, recruitment rarely occurs at a single rate during the course of a trial. Especially in cases of rare diseases, finding eligible patients who are willing to enroll can be much more difficult near the end of the trial than in the middle of it. When this happens, we would not expect the change in outcome attributed to the time trend in a trial with n=128 patients to be the same between patients 30 and 31 versus 120 and 121, when the former patients have a gap between recruitment of 2 weeks and the latter patients have a gap between recruitment of 8 weeks.

If time trends are a concern, the rate at which patients are enrolled and the time at which their responses are measured may also affect analysts' interpretation of the results. If patients are enrolled uniformly across the course of a trial, then Tamm's model from Equation 4.5 is adequate; otherwise, recruitment time must be carefully modeled. We divide our simulation of a clinical trial into three distinct time periods: rampup, steady recruitment, and slowing - or plateauing - recruitment. The rampup period is commonly seen in clinical trials, as sites prepare the logistics of beginning protocols at their respective locations. Many clinical trials have sites formally begin recruitment on different dates, also contributing to a "rampup" period that is marked by slower patient enrollment than most of the remainder of the trial. After all sites have joined and the trial has picked up some momentum, patient enrollment enters a "steady" period where a good number of patients are enrolled each month. Nearing an end of a clinical trial, the third period is often marked by a plateauing recruitment rate, often attributed to difficulty in finding interested or eligible subjects who are willing to enroll.

Since total sample size and recruitment rates alter the length of total recruitment time, recruitment rates are set as percentage of target sample size per month, and the three different stages are defined by the percent time they take of the total recruitment time. Figure 4.5 depicts enrollment of n = 128 patients. The rampup period enrolls 2% of the target sample size per month for the first 15% of the total recruitment time, followed by a steady enrollment period that enrolls 5% of total sample size for the next 60% of recruitment time. The trial ends with slowing recruitment of 2.5 % of the target sample size. Figure 4.6 shows a more aggressive recruitment pattern, enrolling 2%, 20%, and 5% of the target sample size each month during the rampup, steady, and slow periods, respectively.

The components of a design that can be affected by chronological bias are bias, relative bias, Type I error, and power. In this work we call the *naive* analysis the one that does not adjust for an existing time trend in a linear model assessing treatment effect. On the other hand, the *adjusted* analysis includes in the linear model assessing treatment effect a covariate for patient number (j, j_E, or j_C) in the case of uniform enrollment and the time trend affecting both groups, or a covariate for measurement time (m_j) in the case of varying rates of enrollment throughout the trial. The naive analysis yields naive bias, relative bias, Type I error, and power. The adjusted analysis yields adjusted versions of these design components.

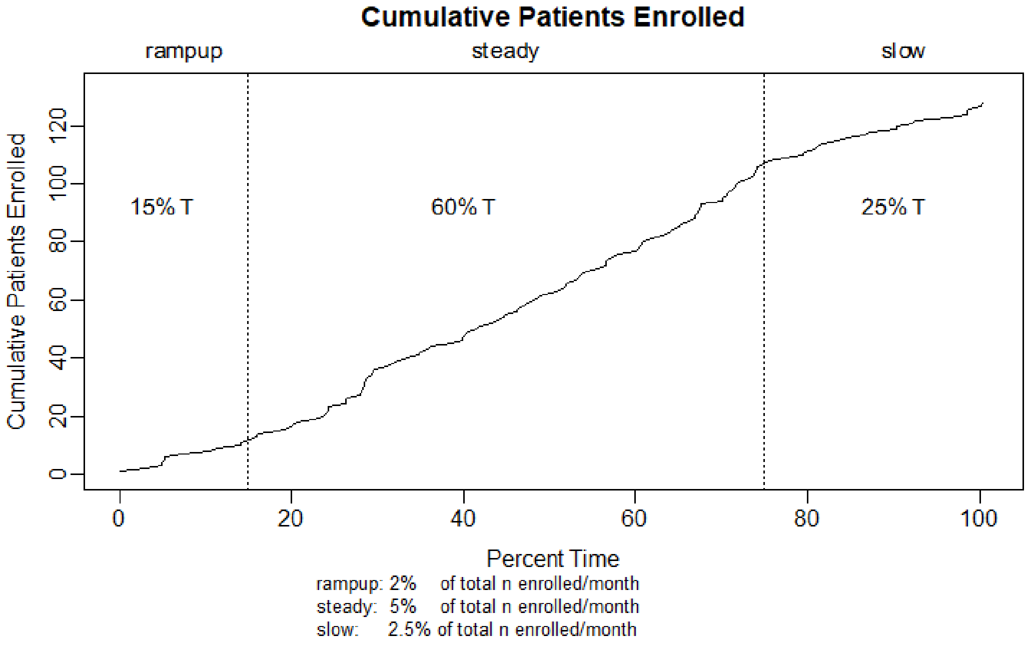


Figure 4.5: Cumulative Patients Enrolled, with 2%, 5%, and 2.5 % of total subjects (n=128) enrolled per month in the rampup, steady, and slow periods, respectively.

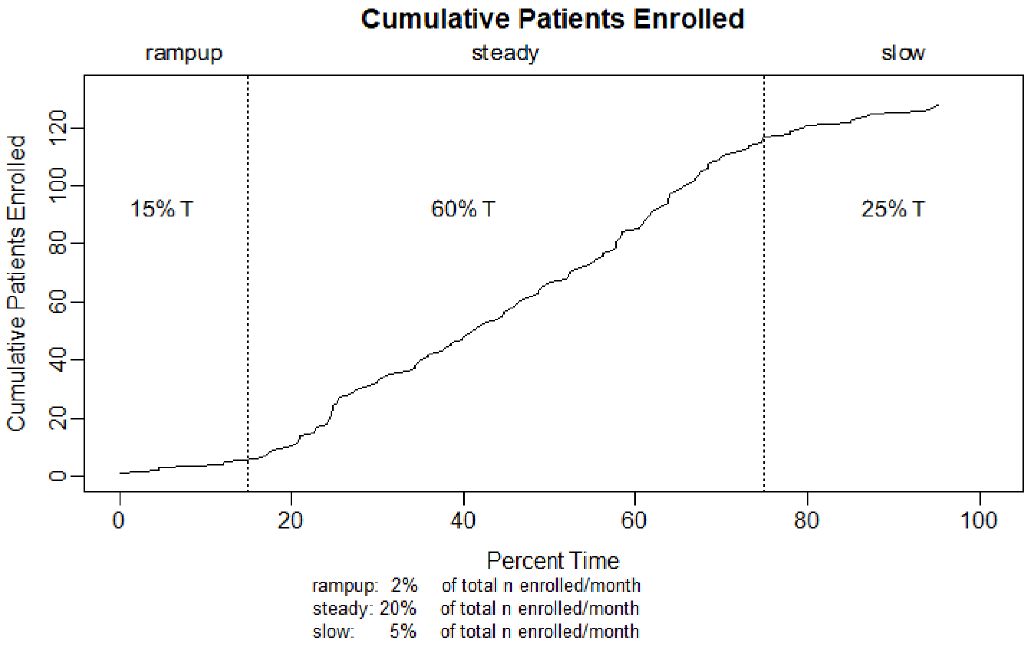


Figure 4.6: Cumulative Patients Enrolled, with 2%, 20%, and 5 % of total subjects (n=128) enrolled per month in the rampup, steady, and slow periods, respectively.

4.1.7 Expected Number of Failures

The Expected Number of Failures of a design is calculated as the average number of failures in the design over a large number of iterations. When the outcome is binary, the outcome is already "success" or "failure", making this calculation straight forward. When the outcome is continuous and considered smaller the better, a failure is defined as any outcome greater than a pre-defined maximum threshold. Let iter = the total number of iterations (the total number of simulated trials), and f_{E_i}, f_{C_i} be the number of failures in the experimental and control arms in iteration i, respectively. Then the

$$\text{Expected Number of Failures} \;=\; \frac{\sum_{i=1}^{\text{iter}} f_{E_i} + f_{C_i}}{\text{iter}}. \tag{4.10}$$

If this characteristic of a design is important to you, consider using a response-adaptive-randomization design (ERADE, DBCD, SMLE, EW1995) that targets the RSIHR allocation. The distribution of observed failures through the simulated trials can be assessed with an individual desirability function.

4.1.8 Expected Total Response

The expected total response is defined as $\bar{Y}_E n_E + \bar{Y}_C n_C$. Two approaches can be taken in assessing this component. The first is to calculate the total response as the average total response across all 10000 iterations by using the average responses across all iterations of a trial and the average treatment group sizes across all iterations of a trial. The second approach evaluates the distribution of observed total responses from each iteration. In Section 4.3.3, both approaches are discussed and the value of the second approach is shown.

When smaller responses are best, a smaller total expected response is desired, and can be targeted with RSIHR allocation (see Section B). When smaller responses are best and correlation is suspected, the target allocation R.corr from Chapter 2 can be utilized.

4.1.9 Bias in the Estimation of the Treatment Effect

We can continue to assess bias and relative bias as defined in Section 2.4. However, the model is now adjusted so that it may adjust for time trends as discussed in Section 4.1.5.

For binary responses, the outcome success ($Y = 1$) or failure ($Y = 0$) is generated using a binomial distribution with success parameter p. Let j represent subject $j \in 1, \cdots, n$, and $h \in c(0, 1)$ represent the null and alternative hypothesis respectively. Let $\boldsymbol{X}, \boldsymbol{\beta} = (\beta_0, \beta_1), \boldsymbol{Z},$ and $\boldsymbol{\beta_{TIME}}$ be defined as in Equations 4.3 and 4.8. The true probability of success is defined as:

$$\boldsymbol{p} = \text{Probability}(Y = 1|j, h) = \frac{exp(\boldsymbol{X\beta} + \boldsymbol{Z\beta_{TIME}})}{1 + exp(\boldsymbol{X\beta} + \boldsymbol{Z\beta_{TIME}})}. \tag{4.11}$$

A simple comparison of the responses is often a first analytical step in clinical trials. In the binary case, this comparison is a difference in success probabilities, and does not take into account other covariates which could be confounding the probability. This is called the naive analysis. On the other hand, the adjusted analysis adjusts for confounders that may be explaining some of the variance of the response. In the context of this work, the true model is known. The naive analysis excludes $\boldsymbol{Z\beta_{TIME}}$ from the analysis of the data. On the other hand, the adjusted analysis includes $\boldsymbol{Z\beta_{TIME}}$. Specifically, for binary responses,

$$\textbf{Naive}\text{: } \hat{\boldsymbol{p}} = \text{Probability}(Y = 1|j, h) = \frac{exp(\boldsymbol{X}\hat{\boldsymbol{\beta}})}{1 + exp(\boldsymbol{X}\hat{\boldsymbol{\beta}})}. \tag{4.12}$$

$$\textbf{Adjusted}\text{: } \hat{\boldsymbol{p}} = \text{ Probability}(Y = 1|j, h) = \frac{exp(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{\boldsymbol{TIME}})}{1 + exp(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{\boldsymbol{TIME}})}. \tag{4.13}$$

For continuous responses, the true response is defined as:

$$E(Y|j,h) = \boldsymbol{X\beta} + \boldsymbol{Z\beta_{TIME}}. \tag{4.14}$$

The naive and adjusted analyses are defined as:

$$\textbf{Naive}\text{: } E(Y|j,h) = \boldsymbol{X\hat{\beta}}, \tag{4.15}$$

$$\textbf{Adjusted}\text{: } E(Y|j,h) = \boldsymbol{X\hat{\beta}} + \boldsymbol{Z\hat{\beta_{TIME}}}. \tag{4.16}$$

Consistent with the simulation setup from Chapter 2.4.1, bias is calculated as $E(\hat{\beta}_1 - \beta_1)$, the expected difference between the estimate of the treatment effect $\hat{\beta}_1$ and the true value of the treatment effect β_1. Relative bias is calculated as $E(\frac{\hat{\beta}_1-\beta_1}{\beta_1}) \times 100$. Since typically we set the true value of the treatment effect β_1 to be 0 under the null hypothesis, the relative bias is only defined under the alternative hypothesis, when $\beta_1 \neq 0$. A distribution of observed bias and relative bias values are evaluated in both naive and adjusted analyses. When a time trend may in fact be responsible for differences in treatment groups, the naive treatment effect estimate is biased and could lead to an incorrect decision of rejecting the null hypothesis.

4.1.10 Type I error & Power of the Design

Two critical characteristics of a design are Type I error and power. In the design stage, the Food and Drug Administration (FDA) places emphasis on controlling Type I error, especially in adaptive designs when multiplicity (multiple looks at the data) is a concern. Type I error is the probability of rejecting the null hypothesis when in fact the null hypothesis is true. On the other hand, the power of the design is the probability of rejecting the null hypothesis given that the alternative is true. In the context of this work, the observed data is used to estimate the treatment effect and its standard error. The null hypothesis is: $H_0 : \beta_1 = 0$, where β_1 is the coefficient for the treatment effect. The alternative hypothesis is: $H_1 : \beta_1 \neq 0$. Formally, Type I error is calculated as

$$P(\text{reject } H_0 \mid H_0 \text{ true}) = P(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} < z_{\alpha/2} \mid \beta_1 = 0). \tag{4.17}$$

It is common for clinical trials to aim to control Type I error at or below 5%.

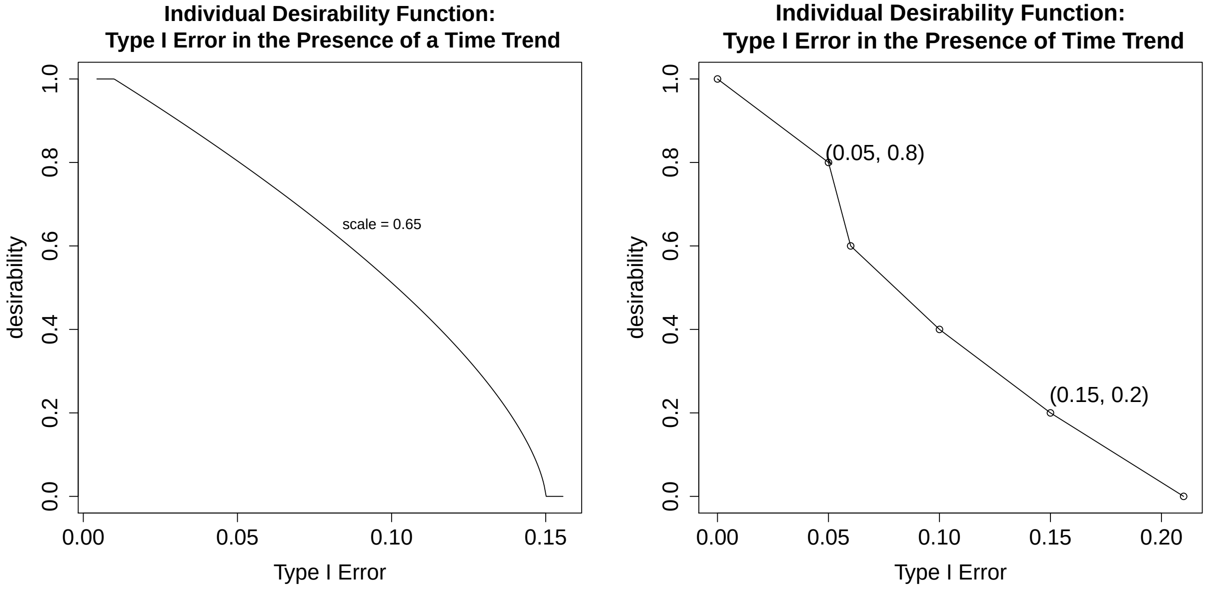
Rosenberger (2016) has commented that power has often become a secondary consideration to the ethical objectives of adaptive designs, yet notes that since accurately reporting no treatment difference is important, power considerations are not to be taken lightly [55, 79]. Power is calculated as

$$P(\text{reject } H_0 \mid H_1 \text{ true}) = P(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} < z_{\alpha/2} \mid \beta_1 \neq 0). \tag{4.18}$$

It is common for clinical trials to aim for high power such as 80% or 90% during the design stage. In this chapter, Type I error and power are calculated in both the naive and adjusted analyses (Equations 4.12, 4.13 for binary responses, Equations B.6, B.7 for continuous responses).

4.2 Determining Scale Parameters and Weights

Two approaches to determining scale parameters are explored. One involves setting scale parameters so that Derringer and Suich's individual desirability functions for each of the components can be created using Equations B.8 through B.10. The second maps specific individual desirability scores to specific values of a component. Individual points are connected linearly so an individual desirability curve is created. Figure B.13 shows an example of each of these methods. This "mapping method" has been used in the literature when a scaling parameter did not accurately reflect opinions of stakeholders or when how the scaling parameter affected the shape of the desirability function was difficult to understand (see Section 3.2.3, [33, 57, 78]). Usage of one method for some characteristics and the alternative method for other characteristics is also possible.



(a) Scale Method

(b) Mapping Method

Figure 4.7: Individual Desirability Function: Type I error.

In Figure B.13a, the scaling method is used with Equation B.9 and scale parameter $r = 0.65$, $L = 0.01$, and $U = 0.15$. This means that any Type I errors below 0.01 will receive an individual desirability score of 1, and any Type I errors above 0.15 will receive an individual desirability score of 0. A Type I error of 0.05 yields an individual desirability of approximately 0.8, highly desirable. Figure B.13b shows a plot based on a mapping of different Type I errors to different desirability scores. It can be seen that a Type I error yields a desirability score of 0.8 (highly desirable), and a Type I error of 0.15 yields a desirability score of 0.2 (unacceptable). The mapping that results in this function is Type I errors of (0, 0.05, 0.06, 0.1, 0.15, 0.21) for individual desirability scores of (1, 0.8, 0.6, 0.4, 0.2, 0), respectively.

It is important that the optimal value of a characteristic and scale parameters accurately reflect an overall consensus on penalizations of different magnitudes of deviations from those target values. It is difficult to use nonlinear least squares to solve for scale parameters to align a characteristic's desirability score with a quantitative gold standard (Section 3.2.9) in the case of evaluating clinical trial design. Although the information matrix is used in finding optimal designs, it reflects the value of certain aspects of a design only. Without a quantitative gold standard that can accurately include all characteristics in consideration, using nonlinear least squares to solve for scale parameters is not an option.

Thus, in the context of evaluation of clinical trial design, we suggest reverting to a method like the one used in Section 3.2.3 - 3.2.5, where a panel of experts is consulted. This is known as the Delphi method. The Delphi method was developed by RAND in the 1950s with the original purpose of forecasting the impact of technology on warfare [70]. Since its development, the Delphi method has maintained popularity in the research community with thousands of citations, showing the value of controlled opinion feedback [24], and performing equally well with other group decision analysis methods such as nominal group technique and social judgment analysis [51, 63].

Utilizing the Delphi method, a panel of statisticians and other clinical trial stakeholders independently provide their opinion on how certain values of clinical trial characteristics or components ought to be scored, and how certain deviations away from target values ought to be penalized. These opinions shape the individual desirability functions which calculate the d_i's of Equation B.11. They should also provide their thoughts on the relative importance - the weights or w_i's - of each characteristic to the overall assessment of the trial. After providing these opinions, these stakeholders should review the feedback in the form of a "group response", or convene to discuss, whereby eventually a consensus opinion ought to be reached on the shapes of individual desirability functions and weights.

4.3 Evaluation of Design Quality: A Framework

The ability of desirability functions to simultaneously assess multiple characteristics or components of a final product were discussed in Section 3. Examples of different and sometimes conflicting clinical trial design characteristics are discussed in Section 4.1. In this Section, a framework is presented so that one can utilize desirability functions to assess the overall weaknesses and strengths of each design under consideration, and quantitatively calculate a standardized score so that designs may be compared objectively. The framework is as follows:

1. Define clinical trial characteristics to be evaluated and specify whether they are Smaller-the-Better, Larger-the-Better, or Nominal-the-Better variables. These characteristics could include, but are not limited to, the components of a design discussed in Section 4.1. Simulate *iter* clinical trials and observe each design's performance in regards to these characteristics.

2. Observe the distribution of values for each characteristic to aid in definition of upper bounds, lower bounds, and target values.

3. Define individual desirability functions for each characteristic using the Delphi method described in Section 4.2. Calculate the individual desirability score for each characteristic and each simulated trial. This means that *iter* scores will be calculated for each characteristic. These individual scores can be examined to understand the strengths and weaknesses of each design.

4. Define weights as described in Section 4.2. Calculate the overall desirability score for each simulated trial using Equation B.11.

5. Observe the distribution of overall desirability scores for each design. Calculate the probability that the overall desirability score is zero for each design:

$$P(D = 0) = \sum_{i=1}^{iter} \mathbb{1}(D_i = 0)/iter.$$

   Use this probability, the mean, standard deviation, and other distributional statistics to help understand the performance of each design within the simulated scenario.

The characteristics evaluated are not limited to those presented in this work. One contribution of the overall desirability score is its ability to incorporate an unlimited number of characteristics. One can incorporate information regarding other statistical characteristics, ethical objectives, or logistic outcomes. Sensitivity analyses may be performed on both shape parameters and weights to observe how design selection may change with altering preferences.

An online tool at `https://priscillakyen.shinyapps.io/DesignEvaluation_beta/` is available for readers to evaluate the quality of different clinical trial designs for two-arm trials.

4.3.1 Application 1: Binomial Responses

Here, we seek to apply the framework of Section 4.3 to an example of design comparison provided in Menon et. al (2015) (Section 10.4.2) [61]. In this example, a two-armed trial expects probability of success in the experimental arm E to be 0.7, and in the control arm C to be 0.4. We denote this with $\boldsymbol{p} = (p_E, p_C) = (0.7, 0.4)$. No time trends are expected. Following the notation of Equations 4.3 and 4.11, $\boldsymbol{\beta} = (\beta_0, \beta_1) = (ln(2/3), ln(3.5)) = (-0.405, 1.253)$.

The authors evaluate non-adaptive designs CRD and PBD, as well as RAR designs DBCD, ERADE, and SMLE targeting RSIHR allocation. As the results showed little difference in the performance of the three RAR designs with respect to treatment failures and power, we will expand upon the designs explored. In this Section, we seek to compare both non-adaptive designs and RAR designs. The designs considered are shown below:

Design	Parameters	Abbreviation
Complete Randomized Design		CRD
Random Block Design Filled With Truncated Binomial	maximum block size = 12	RBD.TBD
Efron's Biased Coin Design	p = 2/3	BCD_2/3
Biased Coin Design with Imbalance Intolerance	p = 2/3, mti = 8	BCDII_2/3
Doubly Biased Coin Design	$\gamma = 2$	
targeting Neyman		DBCD.Neyman
targeting RSIHR		DBCD.RSIHR
Efficient Randomized-Adaptive Design	$\delta = 0.5$	
targeting Neyman		ERADE.Neyman
targeting RSIHR		ERADE.RSIHR

For a brief overview of these designs and the simulation setup used, refer to Chapter 1. For example, for DBCD, refer to Section 1.2. The components of the design evaluated are those discussed in Section 4.1. Menon uses a sample size of $n = 120$ to achieve 92% power of the Wald test under equal allocation. We will utilize a sample size of $n = 106$ to achieve 90% power and maintain Type I error at 0.05. The null hypothesis is $H_0 : \beta_1 = 0$ and the alternative hypothesis is $H_1 : \beta_1 = 1.253$. We report the performance of each design with respect to these components, and discuss individual desirability functions and weighting which result in the ultimate selection of a design.

Treatment Group Size Imbalance

The first component considered is treatment group size imbalance, defined as the difference in group sample size: $n_E - n_C$. Table 4.1 exhibits the performance of the designs considered with respect to this design component.

	Patients in E (mean)	Patients in E (sd)	Proportion in E	Average Treatment Group Imbalance
CRD	52.98	5.20	0.50	-0.04
RBD.TBD	52.98	0.79	0.50	-0.03
BCD_2/3	53.00	1.05	0.50	-0.00
BCDII_2/3	53.04	3.44	0.50	0.08
DBCD.Neyman	58.17	3.33	0.55	10.35
DBCD.RSIHR	60.48	3.82	0.57	14.96
ERADE.Neyman	58.02	2.19	0.55	10.04
ERADE.RSIHR	60.34	2.84	0.57	14.67

Table 4.1: Binary example: treatment group characteristics and size imbalance $n_E - n_C$ under $H_1 : \beta_1 = 1.253$ over 10,000 simulated trials, $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

The designs considered allocated between 41-57% of the predetermined sample size $n = 106$ to the experimental arm E. CRD and the forced balance procedures represent the non-adaptive designs evaluated, and all averaged equal allocation to the two arms over 10,000 simulated trials. The RAR designs targeting Neyman and RSIHR placed more subjects in the better-performing experimental arm. While DBCD and ERADE designs targeting Neyman and RSIHR allocations resulted in similar average number of subjects assigned to the experimental arm, note that ERADE assigned patients to the experimental arm with less variance, with a standard deviation of 2.19 and 2.84 across 10,000 simulated trials for Neyman and RSIHR targets, respectively, compared with DBCD's standard deviation for Neyman and RSIHR targets of 3.33 and 3.82, respectively.

To shape the individual desirability function for treatment group size imbalance, one might consider that the experimental arm has a higher rate of success, so negative treatment group imbalances ($n_E - n_C < 0$) ought to be penalized more steeply. Since the treatment group size imbalance values range from negative to positive, and an imbalance of 0 is considered best, we will treat this component as a Nominal-the-Better (NTB) variable with a target value of 0. The treatment imbalance values of (-35, -25, -15, -8, -3, 0, 8, 12, 15, 22, 30) are mapped to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1, 0.8, 0.6, 0.4, 0.2, 0). This function is used to compute an individual desirability score for each value of treatment group size imbalance yielded from each of 10,000 simulated trials. Table 4.2 summarizes the distribution of scores for treatment group size imbalance. Figure 4.8 plots the individual desirability function for treatment group size imbalance and other assessed design characteristics at the completion of examining all design components. The distribution of scores varies greatly depending on the design. RBD.TBD and biased coin designs BCD_2/3 and BCDII_2/3 never scores below 0.48. The other designs assessed do have treatment imbalance more extreme than -35 or +30, resulting in minimum scores of 0. RBD.TBD and BCD_2/3 have median

	Individual Desirability Scores for Treatment Group Size Imbalance					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.000	0.486	0.700	0.683	0.867	1.000
RBD.TBD	0.680	0.900	1.000	0.954	1.000	1.000
BCD_2/3	0.486	0.867	0.950	0.942	1.000	1.000
BCDII_2/3	0.600	0.680	0.800	0.744	0.850	1.000
DBCD.Neyman	0.000	0.467	0.700	0.629	0.850	1.000
DBCD.RSIHR	0.000	0.257	0.467	0.474	0.700	1.000
ERADE.Neyman	0.000	0.600	0.700	0.669	0.800	1.000
ERADE.RSIHR	0.000	0.314	0.467	0.480	0.700	1.000

Table 4.2: Binary example: summary of individual desirability scores for treatment group size imbalance, $n_E - n_C$ under $H_1 : \beta_1 = 1.253$, for various designs, $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

scores of 1 and 0.95, respectively, which is unsurprising as they are forced balance designs. DBCD.RSIHR and ERADE.RSIHR have the lowest median scores of 0.467, DBCD.RSIHR has the lowest mean score of 0.474. All designs assessed were able to have at least one simulated trial observe a treatment group size imbalance of 0, which resulted in the individual desirability score of 1.

Accidental Bias

The accidental bias factor estimates over 10,000 simulated trials is shown for each of the evaluated designs.

	Accidental Bias Factor Estimates		
	Min	Mean	Max
CRD	0.107	0.110	0.146
RBD.TBD	0.132	0.132	0.133
BCD_2/3	0.124	0.124	0.127
BCDII_2/3	1.178	1.188	1.191
DBCD.Neyman	0.110	0.113	0.684
DBCD.RSIHR	0.109	0.115	0.415
ERADE.Neyman	0.123	0.126	0.890
ERADE.RSIHR	0.115	0.120	0.361

Table 4.3: Binary example: accidental bias factor estimates under $H_1 : \beta_1 = 1.253$ over 10,000 simulated trials, $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

The average accidental bias factor estimate ranges from 0.110 to 1.188. Recall that larger accidental bias factor estimates indicate more substantial bias in the estimate of the treatment effect should other significant confounders be excluded from the model. The amount of bias of the treatment effect estimate is the accidental bias factor multiplied by the squared coefficient of the confounding variable (Section 4.1.2). Not surprisingly, CRD yields the lowest minimum accidental bias factor estimate and the lowest average as well. BCDII_2/3 has an alarmingly high accidental bias factor average of 1.188. Even its lowest accidental bias factor estimate across 10,000 simulated trials is approximately 10 times greater than that of the other designs. Neyman allocations' worst-case accidental bias factor estimates in the simulations are also substantially higher than those of other designs at 0.684 and 0.890 for DBCD and ERADE designs, respectively.

Looking at the distribution of accidental bias factor estimates, the individual desirability function is constructed by mapping accidental bias factor estimate values of $(1, 0.2, 0.13, 0.12, 0.11, 0.1)$ to individual desirability scores $(0, 0.2, 0.4, 0.6, 0.8, 1)$.

	Individual Desirability Scores for Accidental Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.354	0.797	0.828	0.810	0.847	0.853
RBD.TBD	0.392	0.394	0.394	0.394	0.394	0.394
BCD_2/3	0.457	0.520	0.520	0.520	0.522	0.522
BCDII_2/3	0.000	0.000	0.000	0.000	0.000	0.000
DBCD.Neyman	0.079	0.714	0.754	0.732	0.779	0.793
DBCD.RSIHR	0.146	0.654	0.740	0.704	0.779	0.818
ERADE.Neyman	0.028	0.473	0.493	0.485	0.509	0.537
ERADE.RSIHR	0.160	0.564	0.620	0.598	0.661	0.702

Table 4.4: Binary example: summary of individual desirability scores for accidental bias factor estimates under $H_1 : \beta_1 = 1.253$ for various designs, $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

The individual desirability score function computes a distribution of individual desirability scores for each of the accidental bias factor estimates across 10,000 simulated trials. Table 4.4 displays the summary statistics of the resulting scores. The designs have a large range of scores for this component. For example, CRD has the highest maximum individual desirability score of 0.853, and its median is not much lower at 0.810. However, note that RBD.TBD and BCD_2/3 have better worst-case scenarios for this component, with minimum scores of 0.392 and 0.457, respectively, compared with CRD's worst-case score of 0.354. The spread of the distribution and varying performance depending on one's preference for hedging against worst-case scenarios or taking the average performance shows value in studying the entire distribution of estimates rather than just the mean. The high accidental bias factor estimates greater than 1 for BCDII_2/3 result in consistent individual desirability scores of 0. This means that any overall desirability function that gives any positive weight to accidental bias will automatically give BCDII_2/3 an overall score of 0. Second to CRD, DBCD targeting Neyman allocation performs well with an average score of 0.732 across all simulated trials.

Covariate Imbalance

Recall the three types of covariates discussed in Section 4.1.3: C1 is a standard normal variable, C2 represents a covariate that changes linearly over time, and C3 represents an autocorrelated variable. Table 4.5 displays the probabilities of covariate imbalance as estimated by the proportion of 10,000 simulated trials having covariate imbalance exceeding 0.3 for these three covariate types under both the null and alternative hypotheses. We can see that the probability of covariate imbalance for C1 (a standard normal random variable) ranges from 0.122 to 0.131 under the null hypothesis. However, under the alternative, the probability decreases for all designs evaluated; the non-adaptive designs are able to keep balance for C1. RAR designs see an imbalance of C1 exceeding 0.3 around 10% of the time. C2 models a linear trend and is more likely to be

	Under H_0			**Under H_1**		
	C1	C2	C3	C1	C2	C3
CRD	0.122	0.311	0.281	0.000	0.311	0.281
RBD.TBD	0.124	0.142	0.275	0.000	0.000	0.000
BCD_2/3	0.125	0.160	0.255	0.000	0.160	0.255
BCDII_2/3	0.125	0.321	0.209	0.000	0.321	0.209
DBCD.Neyman	0.128	0.245	0.277	0.104	0.259	0.270
DBCD.RSIHR	0.123	0.279	0.265	0.108	0.263	0.276
ERADE.Neyman	0.131	0.174	0.241	0.094	0.199	0.243
ERADE.RSIHR	0.129	0.238	0.246	0.108	0.213	0.254

Table 4.5: Binary example: probability of covariate imbalance under $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = 1.253$ as defined by $|\overline{C}_E - \overline{C}_C| > 0.3, C \in \{C1, C2, C3\}$, $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

imbalanced in CRD and BCDII_2/3 designs under the alternative C3 is an autocorrelated variable and does poorly in most of the designs except for RBD.TBD. Note that RBD.TBD is able to keep covariate balance for all three variable types under the alternative hypothesis. The probabilities (0.30, 0.25, 0.18, 0.15, 0.10, 0) are mapped to individual desirability scores (0, 0.2, 0.4, 0.6, 0.8, 1.0). Since covariate imbalance is more of a concern when there seems to be a significant treatment effect, the individual desirability functions are applied to probability of covariate imbalances under the alternative hypothesis. In this example, our degree of penalization for probabilities of imbalance are consistent regardless of covariate type. In later examples (Section 4.3.2, 4.3.3, 5.2), we demonstrate varying acceptance levels of probability of covariate imbalance for covariates C1, C2, and C3. Table 4.6 shows the resulting individual desirabilities.

	Individual desirability scores for imbalance of 3 covariates		
	C1	C2	C3
CRD	1.000	0.000	0.076
RBD.TBD	1.000	1.000	1.000
BCD_2/3	1.000	0.533	0.179
BCDII_2/3	1.000	0.000	0.316
DBCD.Neyman	0.784	0.165	0.120
DBCD.RSIHR	0.768	0.150	0.095
ERADE.Neyman	0.811	0.347	0.221
ERADE.RSIHR	0.766	0.304	0.184

Table 4.6: Binary example: individual desirability scores for probability of covariate imbalance under the alternative hypothesis $H_1 : \beta_1 = 1.253$ for various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

Since the probability of covariate imbalance is a proportion calculated from the number of simulated trials, each design has only a single individual desirability score value for each of C1, C2, C3, rather than a distribution of individual desirability scores. C1 performs best with our level of tolerance for covariate imbalance, with desirability scores ranging from 0.766 of ERADE.RSIHR to 1.000 for the non-RAR designs. CRD and BCDII_2/3 perform unacceptably with individual desirability scores of 0 with regards to C2, which models a covariate with a linear trend. RBD.TBD performs the strongest with an individual desirability score of 1, which is not surprising due to the blocking nature of the design. Trailing far behind is the second strongest design with respect to balancing C2: BCD_2/3 with an individual desirability score of 0.533. Note CRD's unacceptable score of 0 with regards to C2. RBD.TBD again performs best with C3 with a score of 1. The RAR designs perform poorly, on the other hand, with scores ranging from 0.095 to 0.221, but still score better than CRD with a score of 0.076.

Selection Bias

Table 4.7 displays the selection bias for trials simulated under the alternative hypothesis. Under the null hypothesis, selection bias is less of a concern since no significant treatment difference would render similar probabilities of being assigned to either treatment group. Note that selection bias is always zero for CRD, since the probability of being assigned to either treatment arm is always 1/2. On the other hand, Neyman allocation targeted by DBCD and ERADE have very different performances in regards to selection bias: ERADE is more predictable with a minimum selection bias across all simulated studies of 23.79, compared to DBCD's minimum selection bias of 3.68. RSIHR allocation is similar, with DBCD having lower selection bias values. The larger selection bias of ERADE indicates more extreme probabilities of being assigned to the experimental treatment. Ultimately, as seen in Table 4.1, DBCD and ERADE average the same proportion of patients in the experimental arm, but ERADE's standard deviation for patients assigned to the experimental arm is less than that of DBCD, which confirms the selection bias shown here. The values of (55, 40, 27, 15, 5, 0) are mapped to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1). The individual desirability function is used to calculate a distribution of individual desirability scores from the selection bias values of the designs under evaluation under the alternative hypothesis across 10,000 simulated trials. Table 4.8 summarizes the resulting individual desirability scores for selection bias.

	Selection Bias		
	Min	Mean	Max
CRD	0.00	0.00	0.00
RBD.TBD	6.50	14.68	22.00
BCD_2/3	10.17	13.24	17.33
BCDII_2/3	17.17	25.00	31.67
DBCD.Neyman	3.68	10.08	44.19
DBCD.RSIHR	3.85	9.02	30.48
ERADE.Neyman	23.79	26.41	43.43
ERADE.RSIHR	20.31	25.33	36.31

Table 4.7: Binary example: selection bias under $H_1 : \beta_1 = 1.253$ for various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

CRD has the perfect individual desirability score of 1 for selection bias; other designs have a range of scores from 0.144 to 0.860. While DBCD targeting Neyman allocation a high average score of 0.699, it also has the most extreme worst-case individual desirability score of 0.144. ERADE targeting Neyman

	Individual Desirability Scores for Selection Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	1.000	1.000	1.000	1.000	1.000	1.000
RBD.TBD	0.483	0.583	0.610	0.609	0.630	0.770
BCD_2/3	0.561	0.623	0.637	0.635	0.650	0.697
BCDII_2/3	0.328	0.411	0.433	0.434	0.456	0.564
DBCD.Neyman	0.144	0.664	0.706	0.699	0.741	0.853
DBCD.RSIHR	0.346	0.694	0.729	0.720	0.755	0.846
ERADE.Neyman	0.154	0.401	0.411	0.410	0.420	0.453
ERADE.RSIHR	0.257	0.415	0.427	0.428	0.440	0.512

Table 4.8: Binary example: individual desirability scores for selection bias under $H_1 : \beta_1 = 1.253$ for various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

allocation, being more predictable and having less variation for treatment assignment, has lowest best-case scenario selection bias individual desirability scores of 0.453. From the density plots, ERADE design shows lower selection bias scores than does DBCD, whether they target Neyman or RSIHR allocation and lower variance of scores.

Expected Number of Failures

Table 4.9 displays the expected number of failures under the alternative hypothesis.

	Expected Number of Failures		
	Min	Mean	Max
CRD	28.00	47.72	66.00
RBD.TBD	29.00	47.64	66.00
BCD_2/3	31.00	47.72	68.00
BCDII_2/3	29.00	47.76	66.00
DBCD.Neyman	25.00	46.15	63.00
DBCD.RSIHR	29.00	45.47	63.00
ERADE.Neyman	28.00	46.18	62.00
ERADE.RSIHR	27.00	45.49	65.00

Table 4.9: Binary example: expected number of failures under $H_1 : \beta_1 = 1.253$ for various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

The least number of failures witnessed amongst the designs evaluated across 10,000 simulated trials is 25 failures in DBCD targeting Neyman allocation. On average, RSIHR allocation, whose goal is to minimize the number of failures, indeed has the lowest number of failures whether targeted by DBCD or ERADE, with 45.47 and 45.49 failures for those designs, respectively. Meanwhile, BCDII_2/3 has the largest average number of failures of 47.76. All non-adaptive designs evaluated have an average of about 47.7 failures. Note then, that utilizing DBCD design targeting RSIHR allocation has on average 2.2 failures less than non-RAR designs do. Since 25 is the minimum number of failures observed in the simulation, 25 is mapped to an individual desirability score of 1. The number of failures (70, 58, 46, 37, 30, 25) is mapped to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1). Table 4.10 summarizes the resulting individual desirability scores.

	Individual Desirability Scores for Expected Number of Failures					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.067	0.317	0.367	0.379	0.444	0.880
RBD.TBD	0.067	0.317	0.367	0.380	0.444	0.840
BCD_p	0.033	0.317	0.367	0.378	0.444	0.771
BCD2_p	0.067	0.317	0.367	0.378	0.444	0.840
DBCD.Neyman	0.117	0.350	0.400	0.408	0.467	1.000
DBCD.RSIHR	0.117	0.350	0.422	0.421	0.489	0.840
ERADE.Neyman	0.133	0.350	0.400	0.407	0.467	0.880
ERADE.RSIHR	0.083	0.350	0.422	0.421	0.489	0.920

Table 4.10: Binary example: summary of individual desirability scores for expected number of failures under $H_1 : \beta_1 = 1.253$ for various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

We can see that individual desirability scores range from 0.033 as given by BCD_2/3 design, to 1, as given by DBCD targeting Neyman allocation. On average, the designs evaluated yield individual desirability scores ranging fro 0.327 to 0.421. While both DBCD and ERADE targeting RSIHR designs have the same median and mean individual desirability scores of 0.42 and 0.421 respectively, ERADE proves to be more variable, with lower worst-case individual desirability score of 0.083, compared to DBCD's 0.117. Nonadaptive designs perform similarly.

Bias

Table 4.11 exhibits the performance of the evaluated designs with respect to bias under the null hypothesis $H_0 : \beta_1 = 0$.

	Bias $E(\hat{\beta}_1 - \beta_1)$		
	Min	Mean	Max
CRD	-1.593	-0.003	1.668
RBD.TBD	-1.564	0.001	1.407
BCD_2/3	-1.446	0.001	1.712
BCDII_2/3	-1.763	0.002	1.558
DBCD.Neyman	-17.998	-0.012	1.908
DBCD.RSIHR	-1.674	-0.004	18.332
ERADE.Neyman	-17.733	-0.010	1.617
ERADE.RSIHR	-1.851	0.006	17.926

Table 4.11: Binary example: bias $E(\hat{\beta}_1 - \beta_1)$ for various designs under $H_0 : \beta_1 = 0$ with sample size $n = 106$.

The average bias $E(\hat{\beta}_1 - \beta_1)$ when the null hypothesis $H_0 : \beta_1 = 0$ is true ranges from -0.012 as seen in DBCD design targeting Neyman, to 0.006 as seen in ERADE design targeting RSIHR. The lowest absolute value average bias is 0.001, as seen in RBD.TBD and BCD_2/3. While Neyman allocation resulted in believing the control treatment was better (bias more extreme than -17) in worst-case scenarios, RSIHR allocation resulted in the opposite (bias more extreme than +17) in its worst-case scenarios. The lack of symmetry about the mean for DBCD and ERADE designs targeting Neyman and RSIHR allocations is interesting, as well as the direction of the bias. The bias values of (-5, -2, -0.25, -0.05, -0.01, 0, 0.01, 0.05, 0.25, 1, 5) are mapped to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1, 0.8, 0.6, 0.4, 0.2, 0). Table 4.12 summarizes the resulting individual desirability scores.

	Individual Desirability Scores for Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.167	0.343	0.395	0.437	0.521	1.000
RBD.TBD	0.172	0.343	0.399	0.448	0.509	1.000
BCD_2/3	0.164	0.342	0.395	0.444	0.509	1.000
BCDII_2/3	0.162	0.342	0.395	0.436	0.523	1.000
DBCD.Neyman	0.000	0.342	0.392	0.433	0.519	1.000
DBCD.RSIHR	0.000	0.340	0.393	0.437	0.521	1.000
ERADE.Neyman	0.000	0.343	0.391	0.433	0.521	1.000
ERADE.RSIHR	0.000	0.342	0.393	0.444	0.516	1.000

Table 4.12: Binary example: individual desirability scores for bias $E(\hat{\beta_1} - \beta_1)$ under $H_0 : \beta_1 = 0$ for various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

Table 4.12 shows that the distributions of individual desirability scores for bias under the null hypothesis are similar. The extreme bias values provided by DBCD and ERADE targeting Neyman and RSIHR allocation yielded minimum individual desirability scores for bias of 0. On average, Neyman allocation - targeted by either DBCD or ERADE - had the lowest individual desirability scores for bias of 0.433, slightly lower than the average score of CRD. On the other hand, although RSIHR resulted in occasional individual desirability scores of 0 with regards to bias, the allocation which seeks to minimize treatment failures also averaged equal - when targeted by DBCD - or higher - when targeted by ERADE - individual desirability scores for bias than did CRD. Specifically, DBCD.RSIHR and CRD both averaged a desirability score of 0.437 for bias, and ERADE.RSIHR averaged an individual desirability score of 0.444 for bias. RBD.TBD yielded the highest individual desirability score of 0.448 for bias on average, and did best for worst-case scenarios with a score of 0.172. For all designs, we see the median score is less than the mean score.

Relative Bias

Table 4.13 reports the relative bias of the designs evaluated. Recall that relative bias in this work is defined as $E(\frac{\hat{\beta_1} - \beta_1}{\beta_1}) \times 100$ under the alternative hypothesis.

Table 4.13 shows that on average, the model overestimates the probability of success in the experimental arm, as can be seen by the positive relative bias. The ideal value for relative bias is zero. The true value of β_1 under the alternative hypothesis is $\log(3.5) = 1.253$. Any $\beta_1 > 0$ believes that $p_E > p_C = 0.4$, so relative bias less than -100 indicates the model believes the control arm has a higher probability of success, which is certainly untrue under the alternative. The strongest performer amongst designs evaluated with respect to average relative bias across 10,000 simulated trials is BCDII_2/3, with an average relative bias of

	Relative Bias $E(\frac{\hat{\beta_1}-\beta_1}{\beta_1}) \times 100$		
	Min	Mean	Max
CRD	-116.66	2.23	135.43
RBD.TBD	-112.27	2.42	151.95
BCD_2/3	-117.58	2.25	143.11
BCDII_2/3	-122.96	2.10	143.63
DBCD.Neyman	-205.96	2.96	149.65
DBCD.RSIHR	-106.66	3.20	1484.96
ERADE.Neyman	-1432.54	2.83	223.05
ERADE.RSIHR	-123.85	3.27	186.99

Table 4.13: Binary example: relative bias $E(\frac{\hat{\beta_1}-\beta_1}{\beta_1}) \times 100$ under $H_1 : \beta_1 = 1.253$ for various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$, $\boldsymbol{\beta} = (\beta_0, \beta_1) = (-0.405, 1.253)$ and $n = 106$.

2.10. The weakest performer with respect to average relative bias is ERADE targeting RSIHR allocation, with an average relative bias of 3.27. DBCD targeting RSIHR allocation has extreme values of maximum relative bias of 1484.96 observed in 10,000 simulated trials. On the other hand, Neyman allocation targeted by DBCD and ERADE has extreme minimum relative bias observed of -205.96 and -1432.54, respectively. The strong designs in regards to relative bias when only evaluating min, mean, and max are BCDII_2/3, CRD, BCD_2/3, and Neyman allocation targeted by DBCD. RSIHR allocation targeted by ERADE is a moderate performer, with reasonable min and max of -123.85 and 186.99, respectively, but a relatively high average relative bias of 3.27.

Recall that relative bias less than -100 indicates a belief that the control arm has a higher probability of success, which is known to be untrue. Given this, values below -100 ought to be penalized more heavily. Given this reasoning, we allocate an individual desirability score of 0 when relative bias is less than or equal to the value of -101. Positive relative bias reflects an overestimation of success in the experimental arm, but at least results in a higher probability of correctly rejecting the null hypothesis, and thus should be penalized less. A relative bias of 125 believes that the probability of success in the experimental arm is greater than 0.9, and shall be our maximum acceptable value for positive relative bias. We assign relative bias values of (-101, -75, -50, -25, -10, 0, 10, 25, 55, 90, 125) to individual desirability scores (0, 0.2, 0.4, 0.6, 0.8, 1, 0.8, 0.6, 0.4, 0.2, 0). Table 4.14 summarizes the individual desirability scores calculated from the mapping for relative bias.

The distributions of individual desirability scores are similar. The highest 25th percentile individual desirability score for relative bias is 0.497 of CRD, BCDII_2/3, and ERADE.Neyman. On average, BCD_2/3 performs the strongest for this component with an average individual desirability score of 0.641. Other strong performers include RBD.TBD and CRD, with relatively high mean scores of 0.639, 0.638 respectively, and relative high median scores of 0.630 and 0.637 respectively. Meanwhile, Neyman allocation targeted by DBCD

	Individual Desirability Scores for Relative Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.000	0.497	0.637	0.638	0.785	1.000
RBD.TBD	0.000	0.493	0.630	0.639	0.781	1.000
BCD_2/3	0.000	0.493	0.643	0.641	0.781	1.000
BCDII(2/3)	0.000	0.497	0.644	0.640	0.786	1.000
DBCD.Neyman	0.000	0.494	0.622	0.632	0.786	1.000
DBCD.RSIHR	0.000	0.493	0.628	0.634	0.789	0.998
ERADE.Neyman	0.000	0.497	0.630	0.635	0.789	0.997
ERADE.RSIHR	0.000	0.492	0.633	0.635	0.796	0.997

Table 4.14: Binary example: summary of individual desirability scores for relative bias $E(\frac{\hat{\beta}_1-\beta_1}{\beta_1})\times 100$ under $H_1 : \beta_1 = 1.253$ for various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$, $\boldsymbol{\beta} = (\beta_0, \beta_1) = (-0.405, 1.253)$ and $n = 106$.

performs the weakest on average with an average individual desirability score of 0.632. ERADE.RSIHR has the highest 75th percentile individual desirability score of 0.796, yet never achieves the ideal relative bias of 0 which would have given it an individual desirability score of 1 for this design component.

Type I Error and Power

Table 4.15 displays the performance of the evaluated designs with respect to Type I error and power as calculated as the empirical rejection proportion of the null hypothesis $H_0 : \beta_1 = 0$ under the null and alternative hypotheses, respectively. Because Type I error and power calculates the proportion of rejected null hypotheses across 10,000 simulations, it takes on a single value rather than a distribution of values, so each design will only receive a single individual desirability score for these design components. We see that with a sample size of $n = 106$, Type I error varies from 0.0486 to 0.0540, and power varies from 0.8738 to 0.8862. While ERADE.Neyman and ERADE.RSIHR have the highest power of 0.8809 and 0.8862, respectively, their Type I errors of 0.0518 and 0.0524 are also not controlled as well as those of other designs. Note that Menon's original evaluation of this scenario included $n = 120$ for 92% power, and did not report empirical Type I error as the null scenario was not simulated in that evaluation [61]. RBDTBD, BCD_2/3, and BCDII(2/3) are the three designs able to control Type I error below the nominal $\alpha = 0.05$ level.

	Type I Error	Power
CRD	0.0508	0.8778
RBD.TBD	0.0486	0.8747
BCD_2/3	0.0496	0.8738
BCDII(2/3)	0.0494	0.8783
DBCD.Neyman	0.0540	0.8761
DBCD.RSIHR	0.0513	0.8791
ERADE.Neyman	0.0518	0.8809
ERADE.RSIHR	0.0524	0.8862

Table 4.15: Binary example: Type I error and power for various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$, $\boldsymbol{\beta} = (\beta_0, \beta_1) = (-0.405, 1.253)$, with a sample size of $n = 106$.

DBCD was expected to have power below 90% due to variability of the estimate target allocation proportion $\rho(\hat{\theta_j})_{\text{target}}$, and it is not surprising that ERADE's power is higher than that of DBCD due to lower variability of $\rho(\hat{\theta_j})_{\text{target}}$ (see Sections 1.2.2 and 1.2.3).

The Type I error values of (0.06, 0.0575, 0.0555, 0.0525, 0.05, 0.025) and power values of (0.79, 0.82, 0.84, 0.86, 0.88, 0.90) are given individual desirability scores of(0, 0.2, 0.4, 0.6, 0.8, 1). Table 4.16 exhibits the resulting individual desirability scores.

	Individual Desirability Scores	
	Type I error	Power
CRD	0.736	0.778
RBD.TBD	0.811	0.747
BCD_2/3	0.803	0.738
BCDII(2/3)	0.805	0.783
DBCD.Neyman	0.500	0.761
DBCD.RSIHR	0.696	0.791
ERADE.Neyman	0.656	0.809
ERADE.RSIHR	0.608	0.862

Table 4.16: Binary example: individual desirability scores for Type I error and power calculated from 10,000 simulated studies for various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$, $\boldsymbol{\beta} = (\beta_0, \beta_1) = (-0.405, 1.253)$ and $n = 106$.

Individual desirability scores for Type I error range from 0.5 for DBCD.Neyman to 0.811 for RBD.TBD. Power scores range from 0.738 for BCD_2/3 to 0.862 for RSIHR allocation targeted by ERADE design. Note that none of the power individual desirability scores are 1, since none of the designs yielded a power of 90%.

Overall Desirability Score

So far we have studied the performance of eight different clinical trial designs, each in 10,000 simulated trials. Their results with regards to 11 different design characteristics have shown that some designs are stronger with regards to some characteristics, yet weaker in others, making it no easy task to select a single best design. Now we combine the individual desirability scores to assess the overall performance of each of these designs. The distribution of individual desirability scores was used to calculate a distribution of overall desirability scores. Table 4.17 displays the summary statistics of the overall desirability score D for each of the designs evaluated. Table 4.18 displays the mean individual desirability score for each component assessed and the weights used to calculate the overall desirability scores.

Immediately it is noted that BCDII(2/3) is eliminated from consideration, as its overall desirability score is zero across all 10,000 simulated iterations. RBD.TBD has the highest mean overall desirability score D of 0.639, and the highest median overall D of 0.635. Excluding BCDII(2/3), the weakest performer was ERADE.Neyman, with the lowest average D of 0.544. Although DBCD.RSIHR on average scored higher with a mean overall desirability score of 0.588, it also had a higher probability of 0.038 of scoring 0, compared to ERADE.RSIHR's 0.003.

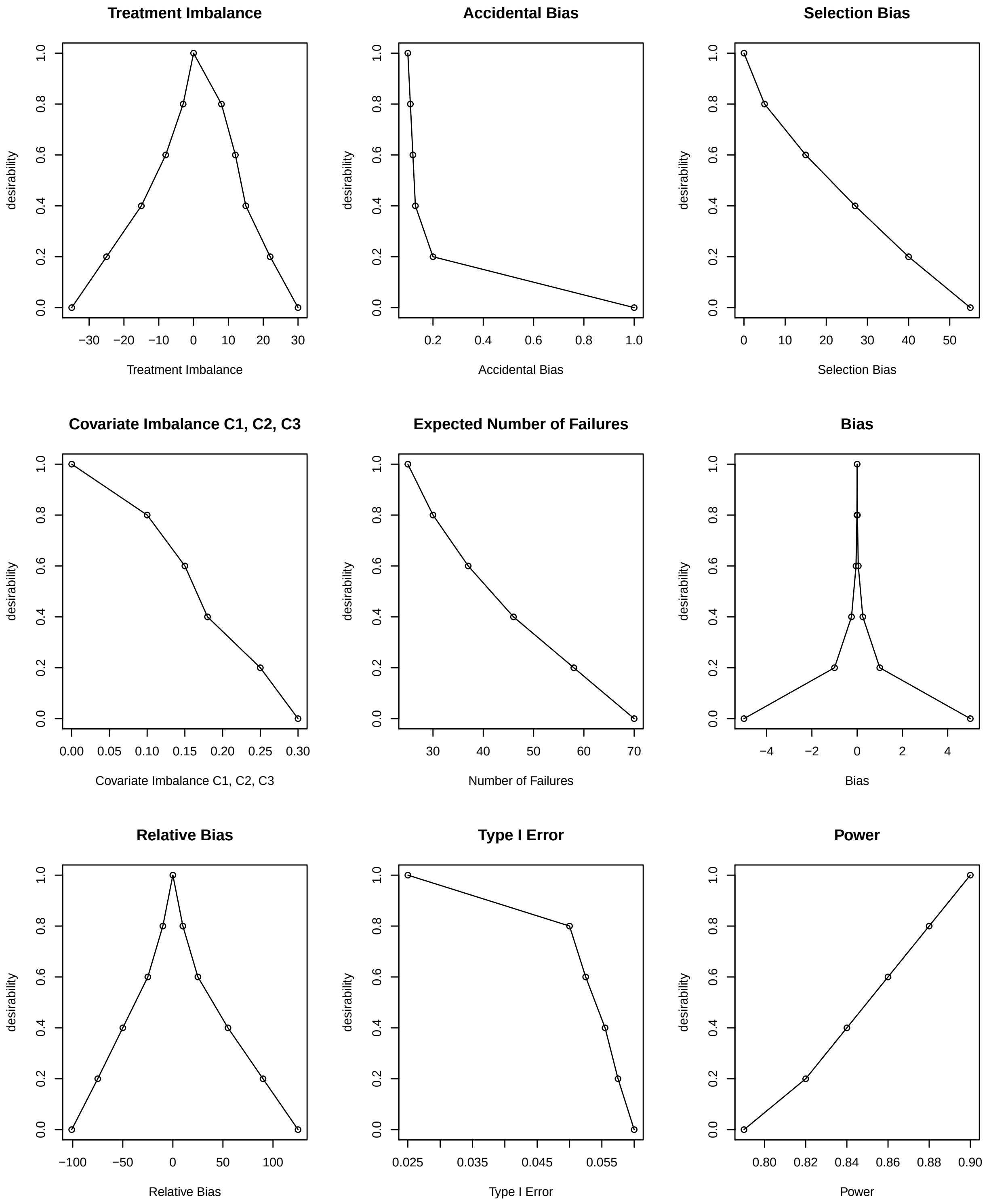


Figure 4.8: Binary example: individual desirability functions.

	Overall Desirability Score D						
	min	q_25	mean	median	q_75	max	Prob(overall D = 0)
CRD	0.000	0.607	0.631	0.625	0.653	0.721	0.004
RBD.TBD	0.000	0.617	0.639	0.635	0.658	0.736	0.001
BCD_2/3	0.000	0.584	0.604	0.601	0.623	0.694	0.001
BCDII(2/3)	0.000	0.000	0.000	0.000	0.000	0.000	1.000
DBCD.Neyman	0.000	0.553	0.576	0.570	0.596	0.672	0.004
DBCD.RSIHR	0.000	0.559	0.588	0.563	0.612	0.705	0.038
ERADE.Neyman	0.000	0.544	0.566	0.560	0.585	0.648	0.003
ERADE.RSIHR	0.000	0.551	0.578	0.564	0.599	0.681	0.014

Table 4.17: Binary example: summary statistics for overall desirability scores of various designs evaluating $\boldsymbol{p} = (0.7, 0.4)$ and $n = 106$.

Table 4.18 recapitulates the means of individual desirability scores for the 11 design components considered, the average overall desirability over 10,000 simulated trials, and the proportion of trials which resulted in an overall desirability of 0. This summary allows us to pinpoint the strengths and weaknesses of each design; nonadaptive designs did well to keep covariate C1 balanced (a standard normal covariate), and also did better controlling Type I error. Although power was weighted more importantly in calculation of the overall score, the weights of covariate imbalance for C1, relative bias, and Type I error were also important, components for which RBD.TBD were strong. We can see that the weakness of the RAR designs is balancing covariate C2, which modeled a linear time trend within the subject covariates (*not* the response), however since we gave this a weight of 0 in this example. The RAR designs also did not do as well in controlling Type I error relative to RBD.TBD. ERADE.RSIHR performed best for bias under the null hypothesis, but the relatively lower weight attributed to the bias component led to a lower overall score.

Overall, Random Block Design with blocks filled by Truncated Binomial Design is the strongest design in this example, given the scale and weight parameters selected. This design yielded the highest mean overall desirability score of 0.639, and has the lowest probability of 0.001 of having a desirability score of 0. It is recommended that trialists test the sensitivity of their weight specifications: for example, changing weight preferences so that expected number of failures has a larger emphasis, with final weights being (0.030, 0.178, 0.119, 0, 0.015, 0, 0.180, 0.060, 0.119, 0.119, 0.179) for the components in the order shown in Table 4.18, resulted in a stronger score provided by complete randomized design.

	CRD	RBD.TBD	BCD_2/3	BCDII(2/3)	DBCD.Neyman	DBCD.RSIHR	ERADE.Neyman	ERADE.RSIHR	weight
Treatment Group Size Imbalance	0.683	0.954	0.942	0.744	0.629	0.474	0.669	0.480	0.030
Expected No. of Failures	0.379	0.380	0.378	0.378	0.408	0.421	0.407	0.421	0.061
Covariate Imbalance									
C1 (N(0,1))	1.000	1.000	1.000	1.000	0.784	0.768	0.811	0.766	0.121
C2 (linear time trend)	0.000	1.000	0.533	0.000	0.165	0.150	0.347	0.304	0.000
C3 (autocorrelated)	0.076	1.000	0.179	0.316	0.120	0.095	0.221	0.184	0.061
Selection Bias	1.000	0.609	0.635	0.434	0.699	0.720	0.410	0.428	0.061
Accidental Bias	0.810	0.394	0.520	0.000	0.732	0.704	0.485	0.598	0.182
Bias	0.437	0.448	0.444	0.436	0.433	0.437	0.433	0.444	0.061
Relative Bias	0.638	0.639	0.641	0.640	0.632	0.634	0.635	0.635	0.121
Type I Error	0.736	0.811	0.803	0.805	0.500	0.696	0.656	0.608	0.121
Power	0.778	0.747	0.738	0.783	0.761	0.791	0.809	0.862	0.182
Overall Desirability D (mean)	**0.631**	**0.639**	**0.604**	**0.000**	**0.576**	**0.588**	**0.566**	**0.578**	
Prob(Overall Desirability D = 0)	**0.004**	**0.001**	**0.001**	**1.000**	**0.004**	**0.038**	**0.003**	**0.014**	

Table 4.18: Binary example: Mean individual desirability scores for 11 design components considered, mean overall Desirability score D, and Probability(D = 0).

4.3.2 Application 2: Clinical Trial Assessing Methotrexate Vs. Placebo in Early Diffuse Scleroderma

In a clinical trial of Methotrexate versus placebo in early diffuse scleroderma patients [68], the primary outcome of interest is total Rodnan skin score measured at twelve months after baseline, with lower values of skin score indicating improvement in the patient. The null hypothesis is that there is no treatment effect, with the expected response in both treatment arms to be 27.5. The alternative hypothesis expects a treatment effect of -6 ($H_1 : \beta_1 = -6$)for the Methotrexate treatment arm, hereby called experimental arm E. Thus, the alternative hypothesis expects the mean response in the placebo group C to be 27.5, and in the Methotrexate group E to be 21.5. Previous literature has indicated higher variance in the treatment group, so $\sigma_A^2 = 219$, and $\sigma_B^2 = 144$. The total sample size to maintain Type I error at the alpha = 0.05 level and power at 80% is $n = 165$ patients.

Subject recruitment is modeled with varying Poisson rates, with the rampup, steady, and slow periods taking 15%, 60%, and 25% of total recruitment time, respectively, and enrolling 2%n, 7%n, and 2.5%n a month, respectively. It is expected that the average level of health of enrolling subjects will be improving over time, and thus the expected outcome is expected to be 1 unit lower at the end of the recruitment period (t = 20 months) relative to the start of the recruitment period, and that this difference is independent of the treatment group. This means that $\lambda = -0.03125$, since the last patient recruited in month 20 will have their measurement taken at 32 months ($32 \times -0.03125 = -1$).

Complete Randomized Design (CRD), Random Block Design Filled with Random Allocation Rule (RBD.RAR) and a maximum block size of 12, PBD with block size of 8 (PBD), Neyman allocation targeted with ERADE design (ERADE.Neyman), and RSIHR allocation targeted with ERADE design (ERADE.RSIHR) are considered. For each design under consideration, 10,000 trials are simulated. The overall performance of these designs with respect to 15 design characteristics are summarized in this section.

Treatment Group Characteristics

Table 4.19 displays a summary of treatment group characteristics. With a total sample size of $n = 165$, it is

	Patients in E (mean)	Patients in E (sd)	Proportion in E
CRD	82.481	6.446	0.500
RBD.RAR	82.499	0.650	0.500
PBD	82.475	0.954	0.500
ERADE.Neyman	88.774	4.403	0.538
ERADE.RSIHR	93.531	4.506	0.567

Table 4.19: Methotrexate trial: treatment group characteristics under $H_1 : \beta_1 = -6$ over 10,000 simulated trials, $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

not surprising that CRD, RBD.RAR, and PBD all end with similar treatment arm sizes, with approximately half of the patients ($n_A = 82.5$) in each of the two arms. The average number of subjects placed in the methotrexate experimental arm begins at 82 subjects (50%) in the CRD design, up to 94 subjects (57%) in the RSIHR allocation targeted by ERADE. The forced balance designs RBD.RAR and PBD, as expected, have lower standard deviations of 0.650 and 0.954, respectively, for patients placed in the experimental arm. The variability of subjects placed in the experimental arm when using RAR designs ERADE.Neyman and ERADE.RSIHR is lower than that of CRD, with standard deviations of 4.403 and 4.506 versus CRD's 6.446. Recall from Section 1.2.3 that the variance of the target allocation within a design influences its power, with larger variances resulting in losses in power.

Treatment Group Size Imbalance

The treatment group size imbalance characteristics of each design are displayed in 4.20. CRD, RBD.RAR, and PBD on average are balanced designs, whilst ERADE.Neyman on average places 13 more subjects, and ERADE.RSIHR on average places 22 more subjects in the methotrexate experimental arm. CRD can place as many as 45 more subjects in the placebo arm than in the experimental arm, more than any other design assessed. Forced balance designs RBD.RAR and PBD both place at most 5 subjects more in either arm. RAR designs ERADE.Neyman and ERADE.RSIHR place at most 19 and 13 more subjects, respectively, in the placebo arm, and at most 53 and 59 more subjects, respectively, in the methotrexate experimental arm.

	Treatment Group Size Imbalance $n_E - n_C$			
	Min	Median	Mean	Max
CRD	-45	-1	-0	51
RBD.RAR	-5	1	-0	5
PBD	-5	-1	-0	5
ERADE.Neyman	-19	13	13	53
ERADE.RSIHR	-13	21	22	59

Table 4.20: Methotrexate trial: summary statistics for treatment group size imbalance, $n_E - n_C$, under $H_1 : \beta_1 = -6$ for designs evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

Although treatment group size balance is typically desired, we can reflect our preference in placing more subjects in the experimental Methotrexate arm by having a small yet positive imbalance mapped to an individual desirability score of 1. Given the small average imbalances in the five designs assessed, a placement of 3 more subjects in the experimental Methotrexate arm than in the control placebo arm will be given an individual desirability score of 1. Deviations away from this value in either direction are given scores less than 1, and thus treatment group size imbalance is considered a nominal-the-better (NTB) design component. Perfect balance as depicted by an imbalance of 0 is also desirable and given a score of 0.8. The remaining individual desirability function is shaped by percentile information. Designs placing more subjects in the placebo arm than in the Methotrexate arm should be penalized more for higher magnitudes of imbalance if the treatment is effective. The (1/7*2, 1/7*3, ..., 1/7*7)th percentiles are mapped to individual desirability scores of 0.8, 0.6, 0.4, 0.2, and 0, respectively. Ultimately, the NTB individual desirability function maps values (0, 0.2, 0.4, 0.6, 0.8, 1, 0.8, 0.6, 0.4, 0.2, 0) to values (-46, -25, -7, -3, 0, 3, 7, 13, 19, 25, 60). Table 4.21 summarizes the resulting individual desirability score distributions from 10,000 simulated trials. RBD.RAR and PBD have the best overall performance with respect to treatment group size imbalance, always having an individual desirability score of at least 0.50. It is noted that CRD and ERADE.RSIHR have similar minimum individual desirability scores, however CRD's performance for this component is generally stronger than that of ERADE.RSIHR, with a mean score of 0.559 compared to ERADE.RSIHR's mean score of 0.359, the lowest of all the considered designs. ERADE.Neyman and CRD have similar distributions, with 25th percentiles at 0.356 and 0.400, 50th percentiles at 0.533 and 0.600, means of 0.559 and 0.582, 75th percentiles both at 0.733, and maximum scores both at 1, respectively.

	Individual Desirability Scores for Treatment Group Size Imbalance					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.010	0.356	0.533	0.559	0.733	1.000
RBD.RAR	0.500	0.733	0.867	0.800	0.867	1.000
PBD	0.500	0.733	0.733	0.795	0.867	1.000
ERADE.Neyman	0.040	0.400	0.600	0.582	0.733	1.000
ERADE.RSIHR	0.006	0.189	0.333	0.359	0.467	1.000

Table 4.21: Methotrexate trial: summary statistics for individual desirability scores for treatment group size imbalance $n_E - n_C$ under $H_1 : \beta_1 = -6$ for various designs, $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

Accidental Bias

Table 4.22 shows Accidental Bias Factor estimates as discussed in Section 4.1.2. The expected bias on the treatment effect is the accidental bias factor estimate × the square of the coefficient of the omitted covariate. It can be seen that CRD, RBD.RAR, and PBD have lower accidental bias factor estimates than those of RAR designs ERADE.Neyman and ERADE.RSIHR. This is not a surprise since the accidental bias factor is inversely proportional to treatment group size imbalance. This results in RBD.RAR and PBD having a uniform distribution of accidental bias factor estimates of 0.053 and 0.049, respectively, since their range of treatment group size imbalance is very small. Neyman allocation as targeted by the ERADE design has average accidental bias factor of 0.059, ranging from 0.058 to 0.072, and RSIHR allocation has the slightly larger range of accidental bias factor ranging from 0.057 to 0.075.

The individual desirability function for accidental bias is shaped by the distribution of observed accidental bias factor estimates in the simulation results. Specifically, the (1/5*5, 1/5*4, 1/5*3, 1/5*2, 1/5*1, 1/5*0)th percentiles drive the decision to map values of (0.076, 0.059, 0.067, 0.053, 0.049, 0.045) to scores of (0, 0.2, 0.4, 0.6, 0.8, 1).

Table 4.23 provides the summary statistics for the individual desirability scores of accidental bias, where

	Accidental Bias Factor Estimates		
	Min	Mean	Max
CRD	0.046	0.047	0.057
RBD.RAR	0.053	0.053	0.053
PBD	0.049	0.049	0.049
ERADE.Neyman	0.058	0.059	0.072
ERADE.RSIHR	0.057	0.059	0.075

Table 4.22: Methotrexate trial: accidental bias factor estimate under $H_1 : \beta_1 = -6$ over 10,000 simulated trials, $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

	Individual Desirability Scores for Accidental Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.405	0.906	0.932	0.916	0.942	0.947
RBD.RAR	0.595	0.600	0.600	0.600	0.600	0.600
PBD	0.796	0.799	0.800	0.800	0.800	0.800
ERADE.Neyman	0.039	0.190	0.200	0.214	0.245	0.262
ERADE.RSIHR	0.012	0.189	0.228	0.239	0.285	0.390

Table 4.23: Methotrexate trial: summary statistics for individual desirability scores for accidental bias factor estimates, $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

the strength of CRD is shown with highest mean score of 0.916 and highest max score of 0.947. CRD's worse-case scenario for this component yields an individual desirability score of 0.405, a score lower than the minimum scores of RBD.RAR and PBD of 0.595 and 0.796, respectively. ERADE.Neyman and ERADE.RSIHR scores for accidental bias range from 0.039-0.262 and 0.012-0.390, respectively, revealing larger impacts on the bias of the treatment effect estimate in the presence of unobserved covariates compared to non RAR designs. Between the two RAR designs, ERADE.RSIHR performs better with respect to accidental bias than does ERADE.Neyman.

Covariate Imbalance

Covariate imbalance is assessed for the three covariates described in Section 4.1.3. The probability of covariate imbalance is defined as $P(|\bar{Z}_A - \bar{Z}_B| > \epsilon$. In this simulation, $\epsilon = 0.3$. Table 4.24 shows the probabilities estimated across 10,000 iterations.

	Under H_0			**Under H_1**		
	C1	C2	C3	C1	C2	C3
CRD	0.051	0.214	0.175	0.000	0.214	0.175
RBD.RAR	0.055	0.058	0.141	0.000	0.000	0.000
PBD	0.052	0.067	0.158	0.000	0.067	0.158
ERADE.Neyman	0.054	0.144	0.159	0.051	0.151	0.160
ERADE.RSIHR	0.056	0.153	0.155	0.049	0.158	0.168

Table 4.24: Methotrexate trial: probability of covariate imbalance, as defined by $|\overline{C}_E - \overline{C}_C| > 0.3, C \in \{C1, C2, C3\}$ under $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = -6$ for various designs evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

It can be seen that the probability of covariate imbalance exceeding 0.3 for C_1 under the null hypothesis is about the same for the five designs evaluated, ranging from 0.051 to 0.056. However, under the alternative hypothesis, it can be seen that C_1 is able to maintain balance when CRD, RBD.RAR, and PBD are utilized. Notice that RBD.RAR and PBD are designs that force balance between two treatment group sizes, and CRD tends towards balance as total sample size n increases. The probability of C_1 differing by more than 0.3 between treatment groups are larger for the RAR designs ERADE.Neyman and ERADE.RSIHR, at 0.051 and 0.049, respectively.

C_2 models a N(0,1) random variable plus a shift that increases with time. It can be seen that forced balance procedures RBD.RAR and PBD perform better with respect to balance of this covariate, with probability of imbalance not exceeding 0.07 under either the null or alternative hypothesis. On the other hand, CRD, ERADE.Neyman, and ERADE.RSIHR, which do not guarantee treatment group size balance, are prone to higher probabilities of covariate imbalance for covariates like C_2. For example, ERADE.RSIHR has the largest probability of imbalance of 0.158 under the alternative hypothesis. This is consistent with expectations: if a time trend is present in a covariate, ensuring that the two treatment group sizes are about the same at several points throughout a trial will lead to more alike values of covariate C_2, whereas trials that are prone to consecutive assignment of subjects to the same treatment arm are at higher risk for imbalance of a covariate subject to a time trend.

C_3 models an autocorrelated variable, meaning that values of the covariate depend on previous responses. The performance of the evaluated designs with regards to C_3 is similar to that of C_2, with the exception of the Permuted Block Design, which had a much higher probability of imbalance of 16% under both null and alternative hypotheses.

For covariate C_1, the (1/4*4, 1/4*3, 1/4*2, 1/4*1, 1/4*1/2)th percentiles of positive probabilities are mapped to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8), and zero probability of imbalance is mapped to an individual desirability score of 1. This equates to the probabilities of (0.170, 0.163, 0.159, 0.127, 0.063, 0) being mapped to scores of (0, 0.2, 0.4, 0.6, 0.8, 1). For covariate C_2 and C_3, quintiles map values of (0.216, 0.170, 0.154, 0.117, 0.053, 0) and (0.176, 0.169, 0.163, 0.159, 0.127, 0), respectively, to scores of (0, 0.2, 0.4, 0.6, 0.8, 1). Table 4.25 displays the resulting individual desirability scores.

RBD.RAR always shows strong performance in the assessed covariate types, with a score of 1 for all three covariates. While CRD is strong in balancing C_1 type covariates, we can see it performs more poorly for covariates C_2 and C_3, with scores of 0.004 and 0.031, respectively, relative to RAR designs. For example, ERADE.Neyman has individual desirability scores of 0.416 and 0.568 for C_2 and C_3, respectively. While this is more desirable than scores of ERADE.RSIHR, note that if we incorporate C_1, ERADE.Neyman overall has a better performance than ERADE.RSIHR, since ERADE.Neyman's scores for C_2 and C_3 are slightly lower

	Individual desirability scores for imbalance of 3 covariates		
	C1	C2	C3
CRD	1.000	0.004	0.031
RBD.RAR	1.000	1.000	1.000
PBD	1.000	0.758	0.606
ERADE.Neyman	0.136	0.416	0.568
ERADE.RSIHR	0.800	0.343	0.245

Table 4.25: Methotrexate trial: individual desirability scores for probability of covariate imbalance under $H_1 : \beta_1 = -6$ for various designs evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

than those of ERADE.RSIHR, but ERADE.Neyman's score for C_1 is significantly lower at 0.136 compared to ERADE.RSIHR's 0.800.

Selection Bias

Table 4.26 shows the performance of the designs evaluated with respect to selection bias. As is expected, selection bias of CRD is 0, since the probability of being assigned to the treatment arm is always 1/2. RBD.RAR has higher selection bias in this scenario since block sizes up to size twelve are allowed, and smaller block sizes yield higher proportion of correct guesses should the block size be known. In reality, the probability of knowing the block size should be very small, since the block size is dynamic and randomly selected. Note that RBD.RAR's ability to have block sizes smaller than PBD's block size of 8 was not sufficient enough to bring down its minimum selection bias of 20.36 below PBD's minimum selection bias of 12.94. ERADE.Neyman and ERADE.RSIHR share similar selection bias values in this case, ranging from 38.30 to 51.18 and 37.86 to 50.15, respectively.

	Selection Bias		
	Min	Mean	Max
CRD	0.00	0.00	0.00
RBD.RAR	20.36	27.36	36.62
PBD	12.94	20.70	33.65
ERADE.Neyman	38.30	41.52	51.18
ERADE.RSIHR	37.86	41.55	50.15

Table 4.26: Methotrexate trial: selection bias for various designs under $H_1 : \beta_1 = -6$ evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

The individual desirability function for selection bias was shaped based off of the (1/5*5, 1/5*4, 1/5*3, 1/5*2, 1/5*1, 1/5*0)th percentiles of selection bias values in the simulated trials, resulting in the values (51.2, 41.4, 37.1, 24.2, 10.4, 0) being mapped to (0, 0.2, 0.4, 0.6, 0.8, 1). Table 4.27 summarizes the resulting individual desirability score distributions for selection bias. CRD has a uniform distribution of individual desirability scores for selection bias at the perfect value of 1, since each subject is always randomized to either arm with equal probabilities. Interestingly, RBD.RAR with maximum block size of 12 has a distribution of scores consistently lower than those of PBD with block size of 8. The 25th ot 75th percentile

	Individual Desirability Scores for Selection Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	1.000	1.000	1.000	1.000	1.000	1.000
RBD.RAR	0.408	0.529	0.553	0.552	0.575	0.656
PBD	0.454	0.629	0.653	0.651	0.675	0.763
ERADE.Neyman	0.000	0.186	0.201	0.206	0.228	0.344
ERADE.RSIHR	0.021	0.185	0.200	0.205	0.227	0.365

Table 4.27: Methotrexate trial: summary statistics for individual desirability scores for selection bias under $H_1 : \beta_1 = -6$ for various designs evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

of ERADE.Neyman and ERADE.RSIHR are very similar from 0.19 to 0.23, but ERADE.RSIHR's minimum and maximum scores of 0.021 and 0.365 are slightly higher than those of ERADE.Neyman's of 0.000 and 0.344, respectively.

Expected Number of Failures

Recall that in the Methotrexate case study, we classify a subject with an outcome of skin score exceeding 31 as having failed. Table 4.28 shows a summary of the number of failures witnessed in the 10,000 simulated trials of each design.

	Expected Number of Failures		
	Min	Mean	Max
CRD	31.00	50.33	71.00
RBD.RAR	30.00	50.35	78.00
PBD	30.00	50.31	72.00
ERADE.Neyman	32.00	49.57	70.00
ERADE.RSIHR	29.00	48.98	73.00

Table 4.28: Methotrexate trial: expected number of failures under the alternative hypothesis for various designs evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

A first comparison of interest notes that ERADE.RSIHR successfully has the lowest average number of failures of 48.98, compared to CRD's average of 50.33. The minimum number of failures is also lowest for ERADE.RSIHR, at 29 failures, compared with CRD's minimum of 31 failures. However, ERADE.RSIHR can potentially yield more failures (73) than all designs assessed, with the exception of RBD.RAR which saw a maximum number of failures of 78.

Percentile statistics again help shape the individual desirability function. The individual desirability function for expected number of failures maps the values of (78, 56, 52, 50, 47, 44) to individual desirability scores of (0, 0.2 0.4, 0.6, 0.8, 1). Note that 78 failures equates to 47% of the total subjects in the trial failing, and 44 failures equates to 27% of the total subjects in the trail failing. 16.5% of the simulated trials had less than 44 failures; these trials would then receive a score of 1 for this component. Table 4.29 summarizes the distribution of individual desirability scores for expected number of failures.

RBD.RAR has the lowest minimum score of 0, meaning at least one of its simulated trials yielded more

	Individual Desirability Scores for Expected Number of Failures					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.064	0.300	0.600	0.570	0.867	1.000
RBD.RAR	0.000	0.300	0.600	0.569	0.867	1.000
PBD	0.055	0.300	0.600	0.570	0.867	1.000
ERADE.Neyman	0.073	0.350	0.600	0.606	0.867	1.000
ERADE.RSIHR	0.045	0.350	0.667	0.636	0.933	1.000

Table 4.29: Methotrexate trial: summary statistics for individual desirability scores for expected number of failures under $H_1 : \beta_1 = -6$ for various designs evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

than 78 failures. PBD and CRD follow next with low minimum scores of 0.055 and 0.064, respectively. ERADE.RSIHR has a minimum score of 0.045. ERADE.Neyman's worst performance in terms of number of failures yielded a score of 0.073. On average, we can see ERADE.RSIHR differentiate itself from the other designs, with the highest mean score of 0.636. ERADE.RSIHR also had the highest median score of 0.667, relative to the other designs, which all had median scores of 0.600. ERADE.RSIHR's 75th percentile score was 0.933, compared with 0.867 of the other designs. On average, non RAR designs scored lower with respect to this component.

Bias

Table 4.30 shows the bias under the null hypothesis across 10,000 simulated trials. Recall that the true difference in success probabilities under the null is 0, and $\boldsymbol{\theta} = (27.5, -6), n = 165$. Table 4.30 shows that non-RAR designs have the same mean bias in both naive and adjusted analyses: CRD, RBD.RAR, and PBD have average biases of 0.172, 0.159, and 0.161, respectively. ERADE.Neyman has an average bias of 0.139 in the analysis not adjusting for time trend, and a slightly lower bias of 0.135 in the analysis adjusting for time trend. ERADE.RSIHR interestingly has higher average bias in the adjusted analysis. The range of bias in the adjusted analysis is broader than that in the naive analysis for all the designs assessed except for PBD. For example, ERADE.RSIHR had a range of bias in the naive analysis of -8.032 to 8.182, and a range of bias in the adjusted analysis of -8.119 to 8.173.

	Naive			Adjusted		
	Min	Mean	Max	Min	Mean	Max
CRD	-6.720	0.172	7.931	-6.898	0.171	8.031
RBD.RAR	-7.189	0.159	7.519	-7.177	0.159	7.571
PBD	-6.891	0.161	7.419	-6.893	0.161	7.419
ERADE.Neyman	-8.196	0.139	9.221	-8.311	0.135	9.213
ERADE.RSIHR	-8.032	0.115	8.182	-8.119	0.117	8.173

Table 4.30: Methotrexate trial: bias of the treatment effect estimate $E(\hat{\beta}_1 - \beta)$ under $H_0 : \beta_1 = 0$ of or adjusted for time trend.

In the definition of the individual desirability function for bias in the naive analysis, the distribution of bias had fatter tails for positive values than for negative values, and the largest bias of 9.221 was considered undesirable. This led to the (1/10*1, 1/10*2, 1/10*3, 1/10*4, 1/10*5)th percentiles of bias values less than 3 mapping to (0, 0.2, 0.4, 0.6, 0.8), the bias of 0 to a score of 1, the (1/10*6, 1/10*7, 1/10*8, 1/10*9)th percentiles of bias values less than 3 mapped to scores of (0.8, 0.6, 0.4, 0.2), and the bias value of 3 mapped to a score of 0. This led to the values (-2.613, -1.710, -1.078, -0.561, -0.066, 0, 0.410, 0.900, 1.426, 2.070, 3.000) mapping to scores of (0, 0.2, 0.4, 0.6, 0.8, 1, 0.8, 0.6, 0.4, 0.2, 0). Table 4.31 reveals the resulting individual desirability scores for bias in the naive analysis.

	Individual Desirability Scores for Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.000	0.093	0.356	0.379	0.628	1.000
RBD.RAR	0.000	0.102	0.360	0.381	0.629	1.000
PBD	0.000	0.100	0.359	0.381	0.630	1.000
ERADE.Neyman	0.000	0.093	0.351	0.376	0.625	1.000
ERADE.RSIHR	0.000	0.087	0.355	0.376	0.624	1.000

Table 4.31: Methotrexate trial: summary statistics for individual desirability scores for bias ($E(\hat{\beta}_1 - \beta)$) under $H_0 : \beta_1 = 0$ without adjusting for time trend under the null hypothesis.

The restriction of having any simulated trial with a bias value greater than 3 led to all designs having a minimum individual desirability score for naive bias of 0. All designs evaluated also had a maximum individual desirability score of 1. ERADE.Neyman had the lowest median score for bias of 0.351. PBD.RAR and PBD had highest median scores of 0.360 and 0.359. It is noted that the performance of the five designs in regards to naive bias under the null hypothesis are similar.

The same rules for defining the individual desirability function for bias in a naive analysis not adjusting for time trend are used to define the individual desirability function for bias in an adjusted analysis incorporating a covariate for time trend. This resulted in bias values of (-2.619, -1.714, -1.082, -0.565, -0.070, 0, 0.406, 0.898, 1.427, 2.073, 3.000) mapping to values of (0, 0.2, 0.4, 0.6, 0.8, 1, 0.8, 0.6, 0.4, 0.2, 0). Table 4.32 summarizes the resulting distribution of individual desirability scores for bias.

	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.000	0.092	0.356	0.379	0.630	1.000
RBD.RAR	0.000	0.103	0.360	0.382	0.632	1.000
PBD	0.000	0.100	0.361	0.381	0.632	1.000
ERADE.Neyman	0.000	0.094	0.349	0.376	0.627	1.000
ERADE.RSIHR	0.000	0.087	0.353	0.376	0.628	1.000

Table 4.32: Methotrexate trial: summary statistics for individual desirability scores for bias ($E(\hat{\beta}_1 - \beta)$) under $H_1 : \beta_1 = 0$ after adjusting for time trends under the null hypothesis.

A comparison between Tables 4.31 and 4.32 reveals that the individual desirability scores for bias in the naive and adjusted analyses are very similar. A look at the 25th and 75th percentiles reveal a very slight increase in scores for bias in adjusted analyses for most designs. Note that a comparison of medians shows that the centers of the distributions of scores for naive and adjusted bias are very similar.

Relative Bias

Table 4.33 exhibits the relative bias across 10,000 simulated trials for the designs evaluated. Recall that a relative bias less than -100 indicates incorrectly that the placebo has the stronger desired treatment effect than the experimental arm. It can be seen that Neyman allocation has the most extreme relative bias in the wrong direction, believing the placebo to be more effective than the treatment arm with a relative bias of -146 and -147 in the naive and adjusted analyses, respectively. Recall from Section 1.2.3 that variability of the proportion of subjects in each arm and the variability of the target allocation proportion contribute to bias. CRD had the largest variability of proportion of subjects in each arm; we see that it also has the largest maximum bias in both the naive and adjusted analyses of 141 and 140, respectively. The RAR designs ERADE.Neyman and ERADE.RSIHR had the next largest variability of patient allocation; its effects on relative bias can be seen as they have the most extreme values of bias in the negative direction of approximately -140 or more extreme in the naive and adjusted analyses.

	Naive			Adjusted		
	Min	Mean	Max	Min	Mean	Max
CRD	-123.162	-7.196	140.833	-124.756	-7.203	140.303
RBD.RAR	-116.424	-6.902	109.992	-115.534	-6.893	110.190
PBD	-130.309	-7.263	130.245	-131.178	-7.255	130.201
ERADE.Neyman	-145.877	-7.295	116.006	-147.064	-7.120	116.539
ERADE.RSIHR	-139.623	-7.242	127.399	-140.144	-7.261	124.711

Table 4.33: Methotrexate vs Placebo Arms: relative bias $E(\frac{\hat{\beta}_1-\beta}{\beta}) \times 100$ under $H_1 : \beta_1 = -6$ naive of or adjusted for time trend, evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

Since relative bias values less than -100 incorrectly suggest a stronger treatment effect in the placebo arm than in the treatment arm, we automatically penalize any relative bias at -101 or more extreme to individual desirability scores of 0. Using CRD as a gold standard, quintiles of relative bias values observed under CRD less than and greater than 0 help define the individual desirability function for relative bias for observed values less than and greater than 0, respectively. This leads to the relative bias values of (-104, -47, -31, -20, -10, 0, 7, 16, 25, 39, 141) to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1, 0.8, 0.6, 0.4, 0.2, 0) for relative bias in naive analyses not adjusting for time trend. Table 4.34 displays the resulting distribution of individual desirability scores for relative bias in the naive analysis.

On average, CRD scores highest with respect to the naive relative bias component, with an average individual desirability score of 0.514. While its median and 75th percentiles are also the highest, note that the 25th percentile score is led by RBD.RAR with a score of 0.261, exceeding that of CRD's 0.258. RAR design ERADE.Neyman does sufficiently well, having comparable scores with CRD in the lower ends

	Individual Desirability Scores for Relative Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.000	0.258	0.507	0.514	0.753	1.000
RBD.RAR	0.000	0.261	0.494	0.507	0.738	1.000
PBD	0.000	0.254	0.503	0.510	0.745	1.000
ERADE.Neyman	0.000	0.256	0.504	0.510	0.746	1.000
ERADE.RSIHR	0.000	0.252	0.500	0.509	0.748	1.000

Table 4.34: Methotrexate trial: summary statistics for individual desirability scores for relative bias $E(\frac{\hat{\beta}_1-\beta}{\beta}) \times 100$ under $H_1 : \beta_1 = -6$ for various designs evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

of the distribution. We can see the distributions begin to separate when looking at the 75th percentile scores: CRD pulling ahead with a score of 0.753 and ERADE.Neyman trailing behind with a score of 0.746. ERADE.RSIHR has a fatter lower tail in the distribution, with lower scores pulling down the 25th and 50th percentiles. Interestingly, ERADE.RSIHR's 75th percentile score is 0.748, higher than ERADE.Neyman's score of 0.746.

A similar decision rule defining the individual desirability function for naive relative bias also defines that of the adjusted relative bias. CRD again is used as the reference or gold standard, leading to the adjusted relative bias values of (-101, 47, -31, -20, -10, 0, 8, 16, 25, 40, 140) to be mapped to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1, 0.8, 0.6, 0.4, 0.2, 0). Table 4.35 displays the resulting distributions of individual desirability scores for relative bias in the adjusted analyses.

	Individual Desirability Scores for Relative Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.000	0.260	0.504	0.514	0.755	1.000
RBD.RAR	0.000	0.262	0.495	0.508	0.741	1.000
PBD	0.000	0.254	0.505	0.511	0.745	1.000
ERADE.Neyman	0.000	0.253	0.506	0.511	0.746	1.000
ERADE.RSIHR	0.000	0.254	0.497	0.509	0.752	1.000

Table 4.35: Methotrexate trial: summary statistics for individual desirability scores for adjusted relative bias $E(\frac{\hat{\beta}_1-\beta}{\beta}) \times 100$ under $H_1 : \beta_1 = -6$ for various designs evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

While CRD's individual desirability score still averaged the highest with a score of 0.514, other designs show improvement with a small upward shift in their scores' values. The exception to this is ERADE.RSIHR, whose 25th percentile, mean, and 75th percentile scores increased, yet whose median decreased from 0.500 to 0.497. Overall, the behavior of the scores is similar, and rankings based off relative bias alone do not change.

Type I Error and Power

The proportion of rejected null hypotheses across 10,000 simulated trials is calculated to understand the Type I error and power in analyses both naive to and adjusted for the time trend. Table 4.36 showcases the tradeoff in performance between Type I error and power resulting from both analyses. For example, ERADE.RSIHR produces reduced Type I error (from 0.0494 to 0.0481) when adjusting for the linear time trend, but also produces reduced power (from 0.7778 to 0.7762). Note that in this scenario, non RAR designs are unable to control Type I error below 0.05, and alternatively have higher power than RAR designs. Interestingly, CRD and RBD.RAR have improved Type I error rates in adjusted analyses, yet PBD has an increased Type I error in the adjusted analysis. The RAR designs' lower power relative to non RAR designs is due to the variability of the target allocation (see Section 1.2.3). ERADE.Neyman's Type I error seems robust to omission of a linear time trend covariate, with similar Type I error rates of 0.0482 and 0.0483 in the naive and adjusted analyses. On the other hand, forced balance designs RBD.RAR and PBD have power levels robust to omission of a linear time trend, with power of 0.787 in both the naive and adjusted analyses in RBD.RAR, and power of 0.7873 and 0.7876 in the naive and adjusted analyses in PBD. RBD.RAR and PBD's insensitivity to omitting a time trend covariate with regards to correctly rejecting the null hypothesis show the value of forced balance designs with regards to robustness in designs expecting time trends.

	Type I error (Naive)	Type I Error (Adjusted)	Power (Naive)	Power (Adjusted)
CRD	0.0533	0.0530	0.7886	0.7851
RBD.RAR	0.0521	0.0513	0.7857	0.7857
PBD	0.0510	0.0519	0.7873	0.7876
ERADE.Neyman	0.0482	0.0483	0.7777	0.7758
ERADE.RSIHR	0.0494	0.0481	0.7778	0.7762

Table 4.36: Methotrexate trial: Type I error and power for various designs evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

While constructing the individual desirability function for Type I error, a clinical trial expecting to deliver results to a regulatory agency might wish to automatically penalize any Type I error above 0.05 with a Type I error individual desirability score of 0. In this case, some leniency is provided so that the designs may still be compared relative to each other. Quintiles are used to shape the individual desirability function, resulting in the naive Type I error values of (0.05430, 0.05234, 0.05144, 0.0536, 0.04916, 0.04500) and the adjusted Type I error values of (0.05400, 0.05212, 0.05154, 0.05010, 0.04826, 0.04500) mapping to the individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1).

Quintiles also were used for guidance in shaping the individual desirability function for power, resulting in naive power values of (0.775, 0.777, 0.7825, 0.788, 0.800, 0.900) and adjusted power values of (0.775, 0.776, 0.782, 0.786, 0.800, 0.900) mapping to values of (0, 0.2, 0.4, 0.6, 0.8, 1). The naive and adjusted power individual desirability functions are close with each other in this case study since the time trend affects both groups and is small. We would expect a time trend unique to a single treatment arm to have larger differentiating effects on Type I error and power in naive and adjusted analyses.

Table 4.37 displays the individual desirability scores for Type I error and power in naive and adjusted analyses.

	Individual Desirability Scores			
	Type I Error (Naive)	Type I Error (Adjusted)	Power (Naive)	Power (Adjusted)
CRD	0.102	0.106	0.610	0.555
RBD.RAR	0.253	0.433	0.516	0.585
PBD	0.481	0.276	0.575	0.623
ERADE.Neyman	0.846	0.796	0.225	0.160
ERADE.RSIHR	0.760	0.810	0.229	0.207

Table 4.37: Methotrexate trial: individual desirability scores for Type I Error and power for various designs evaluating $\boldsymbol{\theta}^T = (27.5, -6)$ and $n = 165$.

CRD has the lowest scores of 0.102 and 0.106 for individual desirability scores for naive and adjusted Type I errors. ERADE.Neyman has the highest score in naive analysis and ERADE.RSIHR has the highest score in adjusted analysis. RBD.RAR and PBD have large score differences with and without adjustments for time trends, with RBD.RAR score's increasing from 0.253 to 0.433, and PBD's score decreasing from 0.481 to 0.276. The range of scores is greater for adjusted power (0.160 to 0.623) than for naive power (0.225 to 0.610).

Overall Desirability Score

In order to calculate the overall desirability scores of the five designs assessed for the Methotrexate trial, the relative importance of the 15 assessed characteristics is considered. Table 4.38 displays the normalized selected weights for treatment group size imbalance, expected number of failures, covariate imbalance C1, covariate imbalance C2, covariate imbalance C3, selection bias, accidental bias factor estimate, bias under the null hypothesis, relative bias under the alternative hypothesis, naive and adjusted Type I error, and naive and adjusted power. These are the w_i's in Equation B.11. Also shown is a reiteration of the mean individual desirability scores of the five assessed designs for each design component. These are the d_i's used in Equation B.11, resulting in the mean overall desirability score shown in bold at the bottom of the table. The probability that the overall desirability score is 0 as estimated by the frequency of the overall desirability score equaling 0 in the 10,000 simulated trials is also shown.

With the individual desirability functions of the 15 components as defined in this section, and the weights as defined in Table 4.38, it is clear that non-RAR designs have better quality than RAR designs. While the RAR designs indeed score higher for expected number of failures and Type I error, and score competitively for bias and relative bias components, non-RAR designs also score sufficiently well for expected number of failures, and significantly better for accidental bias. Specifically, accidental bias and expected number of failures are given the largest weight of 0.143 each. While non-RAR designs score approximately 0.570

	CRD	RBD.RAR	PBD	ERADE.Neyman	ERADE.RSIHR	weight
Treatment Group Size Imbalance	0.559	0.800	0.795	0.582	0.359	0.024
Expected No. of Failures	0.570	0.569	0.570	0.606	0.636	0.143
Covariate Imbalance						
C1 (N(0,1))	1.000	1.000	1.000	0.136	0.800	0.095
C2 (linear time trend)	0.004	1.000	0.758	0.416	0.343	0.000
C3 (autocorrelated)	0.031	1.000	0.606	0.568	0.245	0.000
Selection Bias	1.000	0.552	0.651	0.206	0.205	0.048
Accidental Bias	0.916	0.600	0.800	0.214	0.239	0.143
Bias (Naive)	0.379	0.381	0.381	0.376	0.376	0.024
Bias (Adjusted)	0.379	0.382	0.381	0.376	0.376	0.048
Relative Bias (Naive)	0.514	0.507	0.510	0.510	0.509	0.048
Relative Bias (Adjusted)	0.514	0.508	0.511	0.511	0.509	0.095
Type I Error (Naive)	0.102	0.253	0.481	0.846	0.760	0.095
Type I Error (Adjusted)	0.106	0.433	0.276	0.796	0.810	0.095
Power (Naive)	0.610	0.516	0.575	0.225	0.229	0.048
Power (Adjusted)	0.555	0.585	0.623	0.160	0.207	0.095
Overall Desirability D (mean)	**0.418**	**0.480**	**0.519**	**0.329**	**0.397**	
Prob(Overall Desirability D = 0)	**0.187**	**0.174**	**0.176**	**0.186**	**0.188**	

Table 4.38: Methotrexate trial: mean individual desirability scores for 15 considered design characteristics, mean overall desirability score D, and Probability(D=0).

for expected number of failures compared to RAR designs' scores greater than 0.600, non-RAR designs differentiate themselves as higher quality designs with higher scores in accidental bias: CRD scores 0.916, RBD.RAR scores 0.600, and PBD score 0.800, whilst ERADE.Neyman and ERADE.RSIHR both score below 0.300.

PBD scores highest on average with a mean overall desirability score of 0.519. ERADE.Neyman has the lowest average overall desirability score of 0.329. In deciding the design of the best quality, it is important to look not only at the average overall desirability score D, but also the probability of the overall desirability score being 0. RAR designs ERADE.Neyman and ERADE.RSIHR can be eliminated from consideration with this set of individual desirability functions and weights, with scores of 0.39 and 0.397 respectively, lower than those of the non-RAR designs. ERADE.Neyman and ERADE.RSIHR also have relatively high probabilities of overall desirability score being 0, 0.186 and 0.188, respectively. Amongst PBD's 10,000 simulated trials, 17.6% of them resulted in distributions if individual desirability scores that culminated in an overall desirability score of $D = 0$. This is lower than CRD's 18.7%, and just slightly higher than RBD.RAR's 17.4%. However, PBD's overall mean score of 0.519 is substantially higher than RBD.RAR's 0.480, making it an easy choice in design selection.

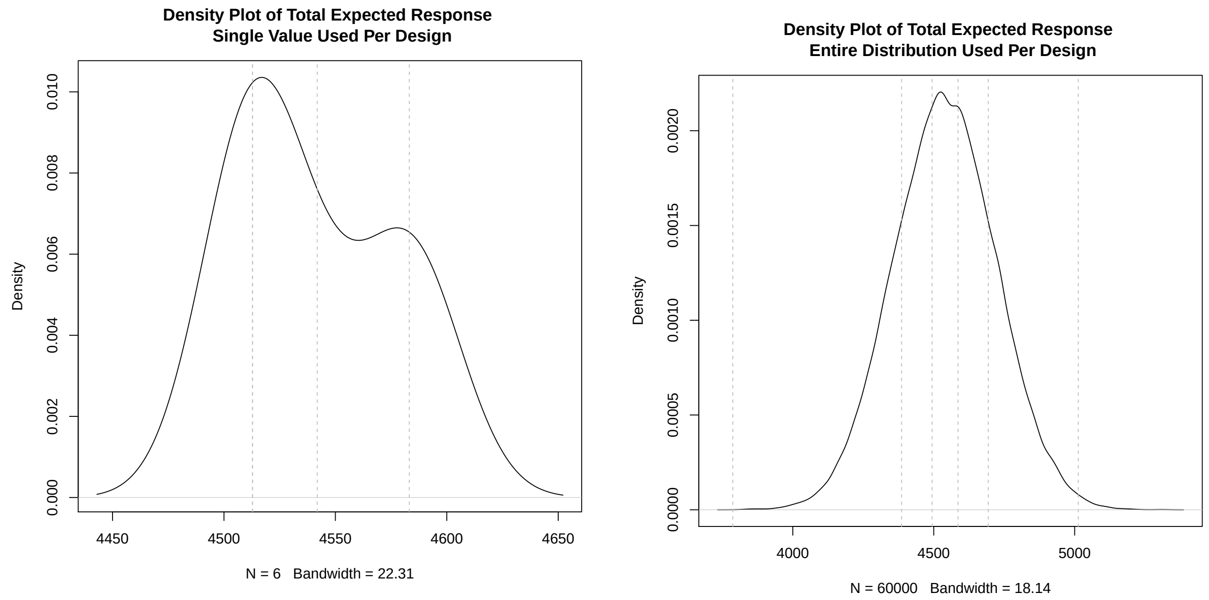
4.3.3 Application 3: Correlated Responses

Let us revisit the Methotrexate vs Placebo example for the scleroderma study from Section 2.4.1[68]. We focus on the hypothetical case where responses between the two treatment arms share a 30% correlation: $\theta = (21.8, 27.5, 219, 144)$, $\rho = 0.30$. The sample size is increased from 165 to 185 to ensure at least 80% power (See Table 2.17). While this seems similar to the example of 4.3.2, we approach this example with no reason to believe a time trend may be present that affects the responses in the two treatment groups. Since no time trend is expected, the naive and adjusted analyses are collapsed into a single analysis, removing design components adjusted bias, adjusted relative bias, adjusted Type I error, and adjusted power. In this section, we incorporate the desire to minimize the total expected response, since lower skin scores are better, by including this design characteristic as one to be assessed in the overall desirability function.

The method of defining individual desirability functions uses percentiles information and is similar to that presented in Section 4.3.2 and will not be repeated here. There are two exceptions: power and total expected response. Since the sample size was increased from 165 to 185 to ensure greater than 80% power, power of 0.8 is mapped to an individual desirability score of 0.8, and power of 0.9 or greater is mapped to an individual desirability score of 1. The range of power witnessed from the simulations was 0.8169 to 0.8334.

Total expected response $\mu_E n_E + \mu_C n_C$ is a newly assessed design component and thus its individual desirability function construction will be discussed more at length here. Two approaches were taken to assess total expected response. The first approach was to calculate $\mu_E n_E + \mu_C n_C$ by using the average number response in each treatment group across all 10,000 simulations and the average number of subjects in each treatment group across all 10,000 simulations. This resulted in each of the assessed designs having a single total expected response value. The individual desirability function could then be built taking the maximum average response of the designs considered when the number of patients placed in the Methotrexate arm is two standard deviations below the mean number of patients placed in the Methotrexate arm and maps this value to an individual desirability score of 0. It then takes the minimum average response of the designs considered when the number of patients placed in the Methotrexate arm is two standard deviations above the mean number of patients placed in the Methotrexate arm, mapping this value to an individual desirability score of 1. The purpose of looking at these tail ends of the distributions is to have a broader range of values included in the calculation of an individual desirability score greater than zero and less than one. Intermediate score values greater than zero and less than one are mapped in the usual manner, using quintiles information from the total expected response of the assessed designs with the average number of patients placed in the Methotrexate arm for each design to inform decisions. This approach of individual desirability function definition led to the values of (4661.18, 4583.17, 4541.90, 4512.83, 4512.79, 4434.26) mapping to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1). Figure 4.9a shows vertical lines at these mapping points; one can deduce that this mapping definition seems to be missing some information, especially with values as close as 4512.83 and 4512.79 mapping to such different individual desirability scores. This mapping led to individual desirability scores ranging from 0.19 to 0.81, with each design's scores spaced almost equally apart.

Identifying the flaws of this first approach, we can improve the individual desirability function by further differentiating the performance of the designs. This is done by looking at the entire distribution of $y_E n_E + y_C n_C$. Since 10,000 trials are simulated for each design, this means that each design will have 10,000 total responses to be scored. When this is done, the quintiles of the total responses are used to help shape the individual desirability function. Specifically, since larger responses are less desirable, we want to penalize positive deviations away from the mean response more than to negative deviations away from the mean. This leads to the midpoint between the fourth and fifth quantiles mapping to a score of 0, followed by the fourth, third, second, first quantiles mapping to scores of (0.2, 0.4, 0.6, 0.8), and the minimum observed total response mapping to a score of 1. This results in the total responses of (5012, 4693, 4586, 4494, 4386, 3787) mapping to values of (0, 0.2, 0.4, 0.6, 0.8, 1). The spread of values in this function will help differentiate the designs in a more meaningful way than does the narrow spread of the first approach. Figure 4.9b displays the density plot of the distributions of total responses across all simulated trials, with vertical lines again at



(a) Density plot of total expected response defined by taking μ_E, μ_C, n_E, and n_C as the average mean response in group E and C, and the average number of subjects in group E and C, across 10,000 simulated studies.

(b) Density plot of total expected response defined by taking the entire distribution of simulated total responses across 10,000 simulated studies.

Figure 4.9: Density plots of total expected response, with gray vertical lines placed at the values mapping to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1).

the mapping points.

Table 4.39 summarizes the distribution of individual desirability scores for expected total responses. The differences in the distributions highlights the value of using the second approach in defining the individual desirability function for this design component.

	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.000	0.192	0.408	0.433	0.657	0.936
PBD	0.000	0.195	0.407	0.435	0.656	0.987
Neyman	0.000	0.261	0.492	0.490	0.736	0.988
RSIHR	0.000	0.312	0.564	0.534	0.798	0.985
RSIHR2	0.000	0.317	0.564	0.536	0.793	1.000
R.corr	0.000	0.315	0.569	0.537	0.800	0.983

Table 4.39: Methotrexate trial with correlated responses: summary of individual desirability scores for total expected response under $H_1 : \beta_1 = -5.7$, where smaller responses are better.

Table 4.39 clearly exhibits the differences in distributions of individual desirability scores for total response; R.corr has the highest median score of 0.569; PBD has the lowest median score of 0.407. On average, CRD has the lowest score of 0.433, and R.corr has the highest score of 0.537. Although R.corr scores highest in the 50th and 75th percentiles and in the mean, its scores are close to those of RSIHR and RSIHR2. Notice that the RAR design DBCD targeting Neyman allocation, which seeks to maximize power, performs very differently than RSIHR, RSIHR2, and R.corr, which all seek to minimize total expected response. For example, on average Neyman allocation scores 0.490 with respect to total expected response, while RSIHR, RSIHR2, and R.corr all score above 0.533.

Individual Desirability Function Definitions

Tables 4.40 and 4.41 display the definitions of the individual desirability functions for the 12 assessed components for nominal-the-better type design components, and for either larger-the-better or smaller-the-better type design components, respectively. Refer to Section 4.3.2 for details on defining the individual desirability functions.

	Nominal-the-Better (NTB) Components		
Individual Desirability Score d	Treatment Group Size Imbalance	Bias	Relative Bias
0	-48.000	-2.436	-101.000
0.2	-16.000	-1.600	-45.396
0.4	-9.000	-0.998	-30.124
0.6	-3.000	-0.491	-18.960
0.8	0.000	-0.028	-9.529
1	3.000	0.000	0.000
0.8	17.000	0.425	7.031
0.6	23.000	0.898	14.637
0.4	27.000	1.422	23.973
0.2	35.000	2.047	37.502
0	76.000	3.000	105.209

Table 4.40: Methotrexate trial with correlated responses: mapping definitions for individual desirability scores for nominal-the-better (NTB) design components.

	Smaller-the-Better (STB) & Larger-the-Better (LTB) Components								
Individual Desirability Score d	Accidental Bias	Imbalance C1	Imbalance C2	Imbalance C3	Selection Bias	Expected No. of Failures	Type I Error	Power	Total Response
0.0	4.214	0.038	0.203	0.157	88.126	88.000	0.054	0.775	5012.822
0.2	0.058	0.037	0.179	0.152	21.734	65.000	0.052	0.777	4693.442
0.4	0.054	0.035	0.176	0.152	17.051	61.000	0.052	0.782	4586.338
0.6	0.041	0.033	0.168	0.152	13.309	59.000	0.051	0.788	4494.315
0.8	0.040	0.032	0.168	0.149	8.964	56.000	0.051	0.800	4385.915
1.0	0.037	0.000	0.051	0.136	0.000	53.000	0.050	0.900	3787.256

Table 4.41: Methotrexate trial with correlated responses: mapping definitions for individual desirability scores for smaller-the-better (STB) and larger-the-better (LTB) design components.

Overall Desirability Score

Using the individual desirability functions defined above, distributions of individual desirability scores were calculated. Table 4.42 shows the mean individual desirability scores of each component. Using these as d_i's, and the weights in the right column as w_i's of Equation B.11, a distribution of overall desirability scores if calculated. Table 4.42 displays the mean and frequency of scores equaling 0.

	CRD	PBD	Neyman	RSIHR	RSIHR2	R.corr	weight
Treatment Group Size Imbalance	0.619	0.799	0.696	0.516	0.517	0.497	0.033
Expected No. of Failures	0.549	0.551	0.592	0.624	0.628	0.626	0.033
Covariate Imbalance							
C1 (N(0,1))	1.000	1.000	0.598	0.211	0.800	0.145	0.131
C2 (linear time trend)	0.200	1.000	0.800	0.400	0.600	0.008	0.000
C3 (autocorrelated)	0.600	1.000	0.400	0.200	0.800	0.042	0.000
Selection Bias	1.000	0.212	0.498	0.494	0.555	0.424	0.016
Accidental Bias	0.443	0.444	0.496	0.539	0.540	0.542	0.197
Bias	0.383	0.383	0.383	0.385	0.384	0.384	0.066
Relative Bias	0.512	0.513	0.512	0.515	0.519	0.515	0.131
Type I Error	0.095	0.200	0.800	0.400	0.909	0.600	0.131
Power	0.857	0.867	0.839	0.837	0.850	0.834	0.066
Total Expected Response	0.433	0.435	0.490	0.534	0.536	0.537	0.197
Overall Desirability D (mean)	**0.372**	**0.407**	**0.488**	**0.409**	**0.548**	**0.408**	
Prob(Overall Desirability D = 0)	**0.175**	**0.169**	**0.170**	**0.169**	**0.162**	**0.170**	

Table 4.42: Methotrexate trial with correlated responses: mean individual desirability scores for 12 considered design characteristics, mean overall desirability score, and probability that overall desirability score is 0.

With the individual desirability functions defined in Tables 4.40 and 4.41, and weights shown in Table 4.42, it is first worth noting that R.corr is not deemed to have the highest overall quality. It successfully performs best with regards to Total Expected Response, with an average individual desirability score of 0.537 for this component, just slightly higher than RSIHR and RHSIR2's scores of 0.534 and 0.536, respectively. R.corr certainly outperforms CRD and PBD in minimizing the total expected response, due to its higher probability of allocating subjects to the experimental Methotrexate arm. In spite of its stronger performance with respect to total expected response, R.corr has an overall quality deemed similar to PBD and RSIHR designs, and loses to RSIHR2, which is deemed to have the highest quality of the designs assessed, with an overall desirability score of 0.548, and probability of overall desirability equaling 0 being the smallest of the designs assessed at 0.162.

Tables 4.42 shows the individual desirability scores that aid our understanding of the strengths and weaknesses of each design; R.corr does well minimizing expected total response and total expected failures, but does substantially poorly in regards to controlling the probability of covariate imbalance for covariate C1. In this regards, R.corr has an individual desirability score of 0.145 for probability of covariate imbalance for C1. RSIHR also fairs poorly with a score of 0.211, whilst the other designs all successfully score nearly 0.6 or higher. R.corr performs equally well with or even better than the other designs in accidental bias, bias, relative bias, and power. While R.corr underperforms relative to RSIHR2 with respect to probability of covariate imbalance, it also underperforms with respect to Type I error, with R.corr having a simulated Type I error of 0.0512, compared to RSIHR2's 0.0505. The weaknesses of R.corr affect its overall quality enough for a trialist to instead turn to RSIHR2 when his preferences are accurately reflected by the individual desirability function definitions and weights described. In a sensitivity analysis, even doubling the weight assigned to total expected response was insufficient to conclude its quality superior to that of RSIHR or RSIHR2.

4.4 Discussion

In this chapter, different components or characteristics of a clinical trial that contribute to its overall quality are discussed. The inclusion of the majority of the components - treatment group size imbalance, accidental bias factor estimate, selection bias (as defined by ρ_{pred} rather than proportion of correct guesses as was Schindler's approach [75]), probability of covariate imbalance, expected number of failures, total expected response, bias, and relative bias - in the scoring of trial quality is a novel use of desirability functions. A framework is provided regarding the decisions behind defining individual desirability functions for each of these components or characteristics, as well as behind the weighting of each of these individual desirability functions in the overall desirability function. Three examples implementing this framework were provided. Upon understanding the framework for assessing design quality using desirability functions, different clinical trial stakeholders should utilize the framework to score design characteristics of interest to them. Other design components that could be considered but are not discussed here include financial cost, time to completion, feasibility of implementation, minimizing the number of subjects assigned to the inferior arm, and probability of stopping early due to early detection of futility or overwhelming efficacy.

The next chapter provides a more in-depth application of the framework introduced here to an HIV clinical trial assessing the efficacy of anti-retroviral treatment in pregnant mothers in reducing the probability of vertical transmission.

Chapter 5

Application to AIDS Clinical Trial

In this chapter, we illustrate how to use desirability functions to select a suitable design for a previously published HIV clinical trial. In 1994, at the first interim analysis of a multicenter clinical trial conducted by the Pediatric AIDS Clinical Trials Group (Protocol 076) evaluating the efficacy of zidovudine (AZT) in the prevention of maternal-infant HIV transmission, the Data and Safety Monitoring Board (DSMB) recommended early termination of the trial in favor of AZT. The DSMB recommended that all patients receiving placebo be provided with the experimental treatment AZT. The trial consequently halted after the first interim analysis.

The approach to design evaluation in this Chapter is similar to that in Chapter 4; however, the application presented here dives into further detail and explores weight sensitivity analyses and analysis of clinical trial designs with different sample sizes. The purpose of the weight sensitivity analysis is twofold: to highlight the importance of accurately defining weights during construction of the overall desirability function, and to remind the reader that certain clinical trial designs have inherent strengths and weaknesses with respect to the evaluated design components

Section 5.1 provides a brief overview of the original clinical trial setup. Section 5.2 explores design selection using the framework presented in Chapter 4 and the trial's original enrolled sample size at the time the trial ended ($n = 477$). Our initial individual desirability functions and weights result in selection of Permuted Block Design with a block size of 8. We find that placing less importance on covariate imbalance leads to a change in selection to DBCD.Baldi. Recall that Baldi allocation has its own weight parameters asking a user to weight ethics and inference; here we have given the Baldi allocation scheme equal weights in ethics and inference. We also find that should we increase our interests in balancing a covariate C1 (a N(0,1) random variable) and in increasing power, our selection shifts to Complete Randomized Design (CRD), the design used in the actual clinical trial. In Section 5.3, we examine how the choice of design changes when each design enrolls only the minimum number of subjects necessary for 90% power. This requires only 190 to 226 subjects to be enrolled for the designs considered. Prior applications have provided each design under evaluation with the same sample size. This new facet can be evaluated quantitatively using a new individual desirability function for sample size needed to attain 90% power. In practice, we can also incorporate the sample size needed into financial considerations for a study. With less subjects available in these simulated studies, each design's weaknesses have little leeway for forgiveness. We see that with all the weight settings evaluated in the prior subsection, RSIHR allocation targeted by the EW1995 RAR design consistently has the highest overall desirability score. Section 5.4 concludes this chapter with a Discussion.

5.1 Original Clinical Trial Setup

A 1994 randomized, double-blind, placebo-controlled trial evaluated zidovudine (AZT) in pregnant women (between 14 and 34 weeks' gestation) with human immunodeficiency virus (HIV) Type I, as defined by having CD4+ T-lymphocyte counts above 200 cells per cubic millimeter. The primary endpoint was maternal-infant transmission of HIV Type I, determined by the 18-month followup of the infant. The target sample size was 636 assessable mother-infant pairs, which was considered able to control Type I error for three interim analyses. Complete randomized design (CRD) was used to randomize mothers to either the AZT or placebo arm [22].

The first interim analysis included data from April 1991 through December 20, 1993. At this time, there were $n_E = 239$ pregnant women placed in the zidovudine experimental arm, and $n_C = 238$ pregnant women placed in the placebo control arm, for a total sample size of $n = 477$ recruited from 59 centers, of which 409 women gave birth to 415 live-born infants. HIV-infection status at 18 months of age was known for 363 infants: 13 infections out of 180 status-known infants from the zidovudine group, and 40 infections out of 183 status-known infants from the placebo group. Forty-six infants were excluded from the analysis because their culture data was not available at the time the interim analysis was performed. The most common reason for data unavailability was due to culture results having yet to be submitted; other reasons included withdrawal before culture was taken, infant was too young, or neonatal death without culture.

The proportions of infants infected were estimated using Kaplan-Meier evaluation of the 363 infants with known HIV-infection status. This yielded estimates of proportion of infants infected as 8.3% in the zidovudine group [95% CI: 3.9-12.8], and 25.5% in the placebo group [95% CI: 8.9-25.5]. The difference in proportions was significant (Z=4.03, two-sided P = 0.00006). The interim stopping boundary required $Z > 3.47$ ($P < 0.0005$). The overwhelming evidence in favor of zidovudine as an antepartum and intrapartum treatment for the mothers, and as a treatment for the newborn, resulted in the DSMB's recommendation to terminate the trial early and to provide zidovudine treatment to all participants of the trial.

5.2 Redesigning the Clinical Trial

We will focus on the trial prior to the first interim analysis and see whether designs other than Complete Randomization (CRD) could have provided further value. We will simulate the trial with simulation setup discussed in Chapter 4. Non-adaptive designs considered are CRD and PBD with block size of 8. Adaptive designs ERADE, DBCD, SMLE, and EW1995 are also considered, targeting Neyman, RSIHR, Urn, and Baldi targets. The Baldi allocation here will place equal weight on ethics and inference. With $p_E = 0.917$ and $p_C = 0.745$, we begin assuming no time trend and have $\boldsymbol{\beta} = (\beta_0, \beta_1) = (1.072121, 1.330146)$.

Treatment Group Characteristics

Table 5.1 summarize treatment group characteristics. The number of patients placed in the AZT experimental arm varies substantially depending on the trial design used, with as little as 183 (38%) patients when using Neyman allocation by EW1995 or ERADE, and as many as 358 (75%) patients when using Urn allocation targeted by ERADE, DBCD, and EW1995. The standard deviation of patient allocation ranges from 0.80 of PBD and 28 of SMLE.Urn. Of the response-adaptive designs, note that those targeting RSIHR allocation have lower standard deviations than those targeting other objectives. Baldi allocation also has low

	Patients in E (mean)	Patients in E (sd)	Proportion in E
CRD	238.48	10.89	0.50
PBD	238.50	0.80	0.50
ERADE.Neyman	183.37	15.94	0.38
ERADE.RSIHR	250.81	2.69	0.53
ERADE.Urn	356.30	20.96	0.75
ERADE.Baldi	327.84	5.37	0.69
DBCD.Neyman	183.75	16.79	0.39
DBCD.RSIHR	250.89	5.70	0.53
DBCD.Urn	358.15	23.02	0.75
DBCD.Baldi	328.80	7.33	0.69
SMLE.Neyman	185.38	21.01	0.39
SMLE.RSIHR	250.79	11.51	0.53
SMLE.Urn	353.60	28.31	0.74
SMLE.Baldi	328.94	11.89	0.69
EW1995.Neyman	183.24	17.63	0.38
EW1995.RSIHR	250.93	7.10	0.53
EW1995.Urn	356.62	25.90	0.75
EW1995.Baldi	329.61	9.58	0.69

Table 5.1: AZT trial: Treatment group characteristics under $H_1 : \beta_1 = 1.330$ over 10,000 simulated trials, $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

patient allocation standard deviation. On the other hand, the Urn allocation sees higher standard deviation exceeding 20. Recall from Section 1.2.3 that the variance of the target allocation within a design influences its power, with larger variances resulting in losses in power.

	Treatment Group Size Imbalance $n_E - n_C$			
	Min	Median	Mean	Max
CRD	-89	-1	-0	83
PBD	-3	1	0	3
ERADE.Neyman	-331	-107	-110	-29
ERADE.RSIHR	5	25	25	45
ERADE.Urn	81	237	236	375
ERADE.Baldi	157	177	179	255
DBCD.Neyman	-323	-107	-109	-21
DBCD.RSIHR	-23	25	25	73
DBCD.Urn	49	241	239	395
DBCD.Baldi	137	179	181	263
SMLE.Neyman	-307	-103	-106	25
SMLE.RSIHR	-59	25	25	115
SMLE.Urn	-31	233	230	403
SMLE.Baldi	97	181	181	285
EW1995.Neyman	-329	-107	-111	-9
EW1995.RSIHR	-27	25	25	81
EW1995.Urn	11	237	236	411
EW1995.Baldi	119	181	182	265

Table 5.2: AZT trial: summary statistics for treatment group size imbalance, $n_E - n_C$, under $H_1 : \beta_1 = 1.330$ over 10,000 simulated trials, $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

Table 5.2 displays summary statistics for treatment group size imbalance $n_E - n_C$. While CRD on average has equal treatment group sizes with a mean treatment group size imbalance of 0, it also allocated as many as 89 less patients in the AZT experimental arm and up to 83 more patients in the AZT experimental arm across 10,000 simulated trials. PBD with a block size of 8 resulted in a maximum absolute value treatment group imbalance of 3. While CRD and PBD on average result in balanced designs, Table 5.2 shows how ERADE, DBCD, SMLE, and EW1995 designs allocate patients for the four allocations Neyman, RSIHR, URN, and Baldi. Note that Neyman allocation on average places 110 *less* patients in the AZT arm, even though the AZT arm has a higher probability of success. Neyman's preference for allocating subjects to an inferior arm is a known characteristic of the allocation for trials with large probabilities of success in both arms. The other three allocations - RSIHR, Urn, and Baldi - all on average place more subjects in the AZT experimental arm. For example, with ERADE design, RSIHR, Urn, and Baldi allocations place on average 25, 236, and 179 more subjects in the AZT experimental arm. Table 5.2 also reveals that ERADE targeting RSIHR, Urn, and Baldi *never* placed more subjects in the control arm during 10,000 simulated trials. While Urn allocation on average placed the most subjects in the AZT experimental arm, note that Baldi allocation places more subjects in the AZT treatment arm than in the control arm, and with less variance than Urn allocation. For example, Urn allocation targeted by DBCD places 239 more subjects in the AZT experimental arm and Baldi allocation targeted by DBCD places 181 more subjects in the AZT experimental arm on average, and the standard deviation of this placement for Urn and Baldi targeted by DBCD is 23.02 vs 7.33, respectively (Table 5.1). Thus although Urn allocation places more subjects in the stronger-performing arm, the lower variability of subject allocation provided by Baldi allocation is something to consider.

While constructing the individual desirability function for treatment group size imbalance, we can decide if a treatment group size imbalance of, say, -10 and +10 should be penalized equally. This may be reasonable if the clinical trial team truly is not sure which treatment arm is superior. However, given prior data, perhaps we can reasonably say we have a belief that the AZT experimental arm has a higher probability of success compared to the placebo control arm, and we would penalize positive imbalances less than negative imbalances. Given this, treatment imbalances of (-100, -50, -20, 0, 33, 60, 90, 150, 200, 250, 300) are mapped to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1, 0.8, 0.6, 0.4, 0.2, 0), respectively. This means that we are scoring designs that place 60 more subjects (about 30%) in the AZT experimental arm than in the placebo arm the highest. The resulting individual desirability score function computes a distribution of individual desirability scores for each of the treatment group size imbalances across 10,000 simulated trials. Table 5.3 shows the summary statistics of the resulting treatment group size imbalance scores.

	Individual Desirability Scores for Treatment Group Size Imbalance					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.044	0.450	0.590	0.573	0.691	0.993
PBD	0.570	0.590	0.606	0.597	0.606	0.618
ERADE.Neyman	0.000	0.000	0.000	0.027	0.044	0.340
ERADE.RSIHR	0.630	0.727	0.752	0.749	0.776	0.889
ERADE.Urn	0.000	0.140	0.252	0.261	0.372	0.860
ERADE.Baldi	0.180	0.460	0.492	0.485	0.516	0.572
DBCD.Neyman	0.000	0.000	0.000	0.033	0.052	0.393
DBCD.RSIHR	0.380	0.703	0.752	0.752	0.800	0.993
DBCD.Urn	0.000	0.116	0.236	0.249	0.364	0.967
DBCD.Baldi	0.148	0.444	0.484	0.478	0.516	0.643
SMLE.Neyman	0.000	0.000	0.000	0.057	0.092	0.752
SMLE.RSIHR	0.164	0.655	0.752	0.742	0.844	0.993
SMLE.Urn	0.000	0.124	0.268	0.286	0.428	0.993
SMLE.Baldi	0.060	0.412	0.476	0.476	0.540	0.777
EW1995.Neyman	0.000	0.000	0.000	0.034	0.052	0.510
EW1995.RSIHR	0.353	0.691	0.752	0.753	0.815	0.993
EW1995.Urn	0.000	0.108	0.252	0.264	0.396	0.993
EW1995.Baldi	0.140	0.428	0.476	0.471	0.524	0.703

Table 5.3: AZT Trial: summary statistics for individual desirability scores for treatment group size imbalance under $H_1 : \beta_1 = 1.330$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

On average, RSIHR allocation targeted by DBCD.RSIHR scores highest in regards to treatment group size imbalance. RSIHR allocation targeted by other designs SMLE, EW1995, and ERADE also score well. Neyman allocation regardless of design has a median and average score close to 0, since it tends to place more subjects in the placebo arm. Even in its best case scenario, Neyman allocation has a maximum score of 0.752 when targeted by SMLE design. Note, however, that high scores occur rarely for SMLE.Neyman, since even its 75th percentile score is 0.092. While Neyman allocation consistently performs poorly with respect to treatment group size imbalance, note that RSIHR allocation performs well even in its worst-case scenarios. Particularly, ERADE.RSIHR's lowest score across 10,000 simulated trials was 0.630, higher than the average score of CRD and PBD of 0.573 and 0.597, respectively. The higher score is a direct result of the decision to shape the individual desirability score such that a positive imbalance receives a higher score than a negative imbalance of the same degree. RSIHR allocation seeks to minimize failures, so its preferential allocation the the AZT experimental arm resulted in it scoring high in this component as well. However, notice CRD's best-case scenario has a maximum score of 0.993, in line with the best-case scenario of RSIHR targeted by SMLE and EW1995. Also note that the individual desirability scores of urn allocation have a long right tail: for example, if we focus on ERADE.Urn, the mean and 75th percentile scores are low at 0.261 and 0.372, but we see that it can also yield a high score of 0.889. The varying shapes of the distributions of the individual desirability scores show the value of not scoring only the mean treatment group size imbalance across 10,000 simulated trials, but each realization of a trial's treatment group size imbalance.

Accidental Bias

Table 5.4 displays the performance of the designs with regards to accidental bias. Accidental bias factor estimates range from 0.006 as shown by RSIHR allocation targeted by SMLE and EW1995, to 0.436 as shown by Urn allocation targeted by EW1995. Recall from Section 4.1.2 that the expected squared bias of the treatment effect estimate due to the omission of a covariate is the accidental bias factor estimate multiplied by the squared effect of the covariate on the outcome (as measured by its coefficient squared). Focusing on the worst-case of an accidental bias factor estimate of 0.436 of EW1995.Urn, this means that for a covariate which results in a 5% decrease in probability of success with a one unit increase of the covariate ($\beta_{omitted} = -0.5275$), the bias on the treatment effect would be $\pm\sqrt{(-0.5275)^2 \times 0.436} = \pm 0.348$. Recall that the true value of $\beta_1 = 1.330146$, so the estimated difference in probability of success attributable to experimental treatment AZT would be $\frac{exp(1.072121+1.330146-0.348)}{1+exp(1.072121+1.330146-0.348)} - 0.745 = 0.886 - 0.745 = 0.141$, rather than the true difference in probability of success of 0.172.

On average, CRD, PBD, DBCD.RSIHR, SMLE.RSIHR, and EW1995.RSIHR had the lowest average accidental bias factor estimates. RSIHR's comparable performance to traditional CRD and PBD indicates that the bias on the treatment effect due to unobserved confounders is comparable between these designs in this trial. Note also that CRD, PBD, and RSIHR allocation has a consistent accidental bias factor estimate

	Accidental Bias Factor Estimates		
	Min	Mean	Max
CRD	0.007	0.007	0.007
PBD	0.007	0.007	0.007
ERADE.Neyman	0.013	0.014	0.047
ERADE.RSIHR	0.008	0.008	0.008
ERADE.Urn	0.019	0.034	0.126
ERADE.Baldi	0.008	0.008	0.012
DBCD.Neyman	0.012	0.014	0.042
DBCD.RSIHR	0.007	0.007	0.007
DBCD.Urn	0.023	0.044	0.231
DBCD.Baldi	0.007	0.008	0.012
SMLE.Neyman	0.018	0.020	0.052
SMLE.RSIHR	0.006	0.007	0.007
SMLE.Urn	0.035	0.066	0.424
SMLE.Baldi	0.007	0.008	0.015
EW1995.Neyman	0.014	0.016	0.052
EW1995.RSIHR	0.006	0.007	0.007
EW1995.Urn	0.029	0.057	0.436
EW1995.Baldi	0.007	0.008	0.012

Table 5.4: AZT trial: accidental bias factor estimate under $H_1 : \beta_1 = 1.330$ over 10,000 simulated trials, $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

throughout the simulated trials, with the minimum accidental bias factor estimate of 0.006 - 0.007 being approximately equal to the maximum accidental bias factor estimate of 0.007. This contrasts with the range of accidental bias factor estimates provided by other RAR designs: for example, urn allocation targeted by DBCD yielded accidental bias factor estimates ranging from 0.023 to 0.231, and Neyman allocation targeted by ERADE from 0.013 to 0.047.

While even the worst-case scenario resulted in an accidental bias of EW1995.Urn of 0.436 seemed to not shift our treatment effect drastically in an example with a one-unit increase in the confounding covariate decreasing the probability of success by 5%, one should still be careful on how they would like to penalize or reward different values of accidental bias factor estimates, since large absolute values of a covariate or its coefficient would result in larger bias in the estimation of the treatment effect. Thus, we may decide that the individual desirability function should heavily penalize even small increases in the accidental bias factor estimate. Referring to Table 5.4 for accidental bias factor estimates, and the overall summary statistics of the combined realizations of accidental bias factor estimates of all designs, the accidental bias factor estimate values are mapped to individual desirability scores. For example, we see that EW1995.Urn had the highest maximum accidental bias factor estimate of 0.436, and thus assign an accidental bias factor estimate of 0.5 to individual desirability score 0. The 75th percentile of all realizations of accidental bias factor estimate was 0.0199; the average was 0.0189. The 25th and 50th percentile values were 0.006834 and 0.008471, respectively. The lowest witnessed accidental bias factor estimate seen was 0.006491, so the value

0.006 is mapped to an individual desirability score of 1. Ultimately, the values (0.5, 0.03, 0.1, 0.008, 0.0068, 0.006) are given scores (0, 0.2, 0.4, 0.6, 0.8, 1). Table 5.5 summarizes the resulting individual desirability scores for accidental bias factors for each design. CRD has reasonably high individual desirability scores for accidental bias, ranging from 0.761 to 0.862. Note the longer left tail of the distribution of scores for CRD, with a minimum score of 0.71, a median of 0.858 greater than the mean of 0.855, and a maximum score of 0.862 just slightly higher than the 75th percentile score of 0.861. Meanwhile, the PBD design consistently yields a score of 0.794 across the 10,000 simulated designs. Neyman allocation regardless of the targeting design has lower scores ranging from 0.2 to 0.5. The average score of Neyman allocation ranges from 0.378 to 0.489, depending on the targeting design. The low worst-case individual desirability scores of 0.064 and 0.076 for urn allocation targeted by EW1995 and SMLE, respectively, warn users of substantial worst-case accidental bias. The urn allocation average had individual desirability scores all below 0.3, further warning of a higher degree of potential accidental bias. On the other hand, the average scores greater than 0.86 by DBCD.RSIHR, SMLE.RSIHR, and EW1995.RSIHR indicate lower accidental bias risk in these designs than in ERADE.RSIHR, with a relatively lower score of 0.594.

	Individual Desirability Scores for Accidental Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.761	0.853	0.858	0.855	0.861	0.862
PBD	0.794	0.794	0.794	0.794	0.794	0.794
ERADE.Neyman	0.205	0.479	0.489	0.484	0.497	0.513
ERADE.RSIHR	0.592	0.593	0.594	0.594	0.594	0.594
ERADE.Urn	0.200	0.211	0.227	0.239	0.256	0.399
ERADE.Baldi	0.526	0.590	0.595	0.594	0.598	0.642
DBCD.Neyman	0.200	0.484	0.495	0.489	0.504	0.518
DBCD.RSIHR	0.797	0.859	0.865	0.864	0.870	0.874
DBCD.Urn	0.200	0.217	0.235	0.243	0.259	0.400
DBCD.Baldi	0.527	0.596	0.618	0.628	0.654	0.780
SMLE.Neyman	0.200	0.366	0.388	0.378	0.403	0.421
SMLE.RSIHR	0.714	0.853	0.868	0.861	0.875	0.877
SMLE.Urn	0.076	0.256	0.286	0.295	0.328	0.399
SMLE.Baldi	0.477	0.584	0.594	0.610	0.622	0.842
EW1995.Neyman	0.200	0.442	0.456	0.449	0.466	0.484
EW1995.RSIHR	0.786	0.859	0.868	0.865	0.874	0.877
EW1995.Urn	0.064	0.236	0.260	0.271	0.294	0.399
EW1995.Baldi	0.525	0.595	0.617	0.635	0.672	0.855

Table 5.5: AZT Trial: summary statistics for individual desirability scores for accidental bias factor estimates under $H_1 : \beta_1 = 1.330$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

Covariate Imbalance

Recall the three types of covariates discussed Section 4.1.3: C1 is a standard normal variable, C2 represents a covariate that changes linearly over time, and C3 represents an autocorrelated variable. Table 5.6 displays the probabilities of covariate imbalance exceeding 0.3 for these three covariates under the null and alternative hypotheses. Note first that the probabilities of covariate imbalance for covariates C1, C2, and C3 are very similar whether or not there is a treatment effect. For example, the probability of imbalance for covariate C1 is 0.001 and 0.000 under the null and alternative hypotheses, respectively, for either CRD or PBD. The estimated probability of imbalance of C2 is 0.031 under CRD under either the null or alternative hypotheses. The insensitivity of probability of covariate imbalance to whether or not there is a treatment effect for CRD and PBD is expected, since the allocation rules of these designs do not depend on the estimate of the treatment effect.

	Under H_0			**Under H_1**		
	C1	C2	C3	C1	C2	C3
CRD	0.001	0.031	0.022	0.000	0.031	0.022
PBD	0.001	0.001	0.016	0.000	0.001	0.016
ERADE.Neyman	0.001	0.010	0.015	0.002	0.100	0.025
ERADE.RSIHR	0.001	0.002	0.010	0.001	0.002	0.011
ERADE.Urn	0.001	0.122	0.030	0.005	0.318	0.062
ERADE.Baldi	0.002	0.007	0.027	0.002	0.014	0.023
DBCD.Neyman	0.001	0.021	0.022	0.001	0.104	0.026
DBCD.RSIHR	0.001	0.010	0.020	0.001	0.012	0.019
DBCD.Urn	0.000	0.136	0.019	0.004	0.282	0.052
DBCD.Baldi	0.002	0.022	0.038	0.002	0.030	0.034
SMLE.Neyman	0.001	0.043	0.021	0.001	0.082	0.025
SMLE.RSIHR	0.001	0.030	0.020	0.000	0.034	0.018
SMLE.Urn	0.000	0.129	0.023	0.005	0.266	0.050
SMLE.Baldi	0.003	0.058	0.034	0.002	0.050	0.032
EW1995.Neyman	0.001	0.027	0.019	0.001	0.108	0.024
EW1995.RSIHR	0.001	0.018	0.020	0.001	0.015	0.022
EW1995.Urn	0.001	0.134	0.021	0.005	0.261	0.054
EW1995.Baldi	0.003	0.039	0.036	0.002	0.044	0.033

Table 5.6: AZT trial: probability of covariate imbalance, as defined by $|\overline{C}_E - \overline{C}_C| > 0.3, C \in \{C1, C2, C3\}$ under $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = 1.330$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

Under the null hypothesis H_0, most designs do well avoiding covariate imbalance for C1, with the proportion of simulated trials resulting in C1 covariate imbalance exceeding 0.3 ranging from 0.000 to 0.003, with Baldi allocation resulting in highest proportion of 0.002 to 0.003. On the other hand, under the alternative hypothesis, we observe a larger range in probabilities of covariate imbalance for C1, with Urn target having an estimated probability of 0.005. On average, the probability of covariate imbalance for C1 under either the null or alternative is acceptably small in this trial.

The probability of imbalance of covariate C2 using CRD and PBD is 0.031 and 0.001, respectively, regardless of whether the trials are simulated under the null or alternative hypotheses. The highest risk for imbalance of C2 is present in urn allocation targeted by DBCD design under the null, with an estimated probability of imbalance of 0.136, and in urn allocation targeted by ERADE design under the alternative, with an estimated probability of imbalance of 0.318. Under the alternative, many of the RAR designs have lower probability of imbalance of C2 than CRD, including ERADE.RSIHR, ERADE.Baldi,and DBCD.Baldi. The value of simulation is evident here, with Neyman and Urn allocations proving less protective against imbalance of a covariate with a linear trend than other target allocations Baldi and RSIHR.

Lastly, the probability of imbalance of covariate C3 using CRD and PBD is 0.022 and 0.016, respectively, regardless of whether the trials are simulated under the null or alternative hypotheses. The highest risk for imbalance of C3 under the null hypothesis is when targeting Baldi utilizing DBCD, with a probability of imbalance of 0.038. ERADE targeting urn allocation under the alternative hypothesis has a probability of imbalance of 0.062. Opposite to the relative performance of these designs with C2, Urn and Baldi target allocations proved to have higher probabilities of imbalance for C3 than Neyman and RSIHR allocations.

While in Chapter 4 we have provided the same individual desirability function for C1, C2, and C3, here in this more applied case study we show the flexibility of the framework by providing a distinct individual desirability function to score each covariate type's probability of imbalance exceeding 0.3. We continue to focus on covariate imbalance scenarios only when there is a treatment effect present (under the alternative hypothesis). Covariate imbalance is an undesirable characteristic, and is thus of the smaller-the-better type, with values of 0 receiving the highest score of 1. Since the proportion of covariate imbalance exceeding 0.3 for covariate C1 ranges from 0 to 0.005, we map a value of 0.006 and greater to an individual desirability score of 0. Specifically, probabilities (0.006, 0.004, 0.001, 0.005, 0.002, 0) are mapped to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1).

For the C2 covariate representing a linear trend, the proportions of imbalance in 10,000 simulated trials ranged from 0 to 0.062. We would like to automatically penalize designs with a probability of C2 covariate imbalance greater than 0.25 to have an individual desirability score of 0 for this design characteristic. While a probability of 0 yields a perfect score of 1, we do see some designs such as PBD and ERADE.RSIHR yielding low probabilities of imbalance of 0.001 and 0.002. After that, we see the probability values jump towards 0.3. Given this, the probabilities (0.25, 0.15, 0.10, 0.0325, 0.0010, 0) are mapped to scores (0, 0.2, 0.4, 0.6, 0.8, 1).

The C3 covariate represents an autocorrelated covariate, where the value of the subject's covariate depends on that of a previous subject. While this is not a concern regarding the response, it is a feasible type of independent variable. The proportions of covariate imbalance for C3 in 10,000 simulated trials ranged from 0 to 0.062. Designs ERADE.RSIHR yield low probabilities of imbalance of 0.011, so we will map the value of 0.01 to a score of 0.80. The probabilities of imbalance of C3 of the other designs seem to range uniformly; the probability values (0.06, 0.045, 0.03, 0.02, 0.01, 0) are mapped to scores (0, 0.2, 0.4, 0.6, 0.8, 1).

Table 5.7 shows the resulting individual desirability scores for probability of covariate imbalance for the discussed three covariates. PBD performs well balancing C1 and C2, and relatively well for C3, with scores

	Individual desirability scores for imbalance of 3 covariates		
	C1	C2	C3
CRD	1.000	0.608	0.560
PBD	1.000	0.799	0.682
ERADE.Neyman	0.327	0.398	0.508
ERADE.RSIHR	0.560	0.793	0.772
ERADE.Urn	0.120	0.000	0.000
ERADE.Baldi	0.360	0.716	0.534
DBCD.Neyman	0.387	0.386	0.472
DBCD.RSIHR	0.440	0.731	0.620
DBCD.Urn	0.160	0.000	0.105
DBCD.Baldi	0.320	0.617	0.341
SMLE.Neyman	0.380	0.453	0.506
SMLE.RSIHR	0.600	0.595	0.632
SMLE.Urn	0.050	0.000	0.132
SMLE.Baldi	0.340	0.550	0.376
EW1995.Neyman	0.387	0.368	0.528
EW1995.RSIHR	0.560	0.710	0.556
EW1995.Urn	0.110	0.000	0.085
EW1995.Baldi	0.307	0.567	0.361

Table 5.7: AZT Trial: individual desirability scores for probability of covariate imbalance under $H_1 : \beta_1 = 1.330$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

of 1, 0.799, and 0.682, respectively. Note that although CRD has a score of 1 in balancing C1, several other adaptive designs are able to outperform it for C2 and C3. For example, ERADE.RSIHR has scores of 0.793 and 772 for C2 and C3, respectively, compared with CRD's 0.608 and 0.560, respectively. Urn and Neyman allocations have low individual desirability scores for all three covariates C1, C2, and C3. For example, urn allocation targeted by EW1995 has undesirable scores of 0.110, 0.000, and 0.085, for C1, C2, and C3 respectively. Recall Neyman allocation tends to place more subjects in the AZT experimental arm, and here we note that it also performs poorly in balancing covariates. Baldi allocation has a score of approximately 0.3, depending on the targeting design, for standard normal covariate C1, and score below 0.4 for C3 for DBCD, SMLE, and EW1995 designs, but does better under ERADE with an individual desirability score of 0.534 for covariate C3. Although Baldi allocation does not do well in avoiding imbalance of C1 and C2, note that it scores above 0.7 for covariate C2, regardless of the targeting design. For example, ERADE.Baldi has a score of 0.716 for C2, higher than the score of CRD, and underperforming only relative to ERADE.RSIHR and PBD designs.

Selection Bias

Table 5.8 displays the performance of the evaluated designs in regards to selection bias. Recall from Section 4.1.4 that third-order selection bias occurs when past subject allocation is known or can be correctly guessed by an investigator, influencing their ability to predict and act on their belief of future allocations based on prior assignments. Thus, the higher the selection bias measure shown in Table 5.8, the more susceptible a trial's results are to selection bias, which influences validity of the treatment effect estimate since certain subjects may be more likely to be assigned to one arm or the other.

	Selection Bias		
	Min	Mean	Max
CRD	0.00	0.00	0.00
PBD	47.47	60.51	79.23
ERADE.Neyman	104.79	116.78	126.16
ERADE.RSIHR	103.12	115.76	121.48
ERADE.Urn	104.80	119.15	133.52
ERADE.Baldi	125.72	139.85	155.79
DBCD.Neyman	8.80	20.27	73.38
DBCD.RSIHR	7.83	16.10	41.87
DBCD.Urn	13.00	37.74	97.01
DBCD.Baldi	71.96	101.50	136.67
SMLE.Neyman	2.65	10.45	61.31
SMLE.RSIHR	1.96	5.98	18.45
SMLE.Urn	6.71	26.81	99.18
SMLE.Baldi	88.48	99.35	121.38
EW1995.Neyman	6.19	15.64	63.31
EW1995.RSIHR	4.95	11.58	30.02
EW1995.Urn	12.59	33.72	221.65
EW1995.Baldi	81.96	101.20	133.83

Table 5.8: AZT trial: selection bias under $H_1 : \beta_1 = 1.330$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

By definition, the selection bias of CRD is 0, since the probability of being allocated to either arm is always 0.5, regardless of the stage of the trial or any prior results. PBD with block size 8 results in an average selection bias of 60.51. Selection bias in PBD is less of a concern if the block size is able to be hidden from the investigator. However, should the investigator know the block size, he may occasionally enroll subjects with an idea of the probability of assignment to the AZT experimental arm. Of the RAR designs, Table 5.8 shows that, save for Urn design targeted by EW1995, the ERADE design has much higher selection bias of over 120 compared to the same target allocations pursued by DBCD, SMLE, or EW1995. ERADE's larger selection bias than other RAR designs is not surprising due to its attainment of the lower bound of the variance of n_E/n, as discussed in Section 1.2.2. Amongst the RAR designs, RSIHR targeted by SMLE had the lowest min, mean, and max selection biases of 1.96, 5.98, and 18.45, respectively. Conversely, EW1995.Urn saw the highest worst-case selection bias of 221.65. When implementing DBCD, RSIHR yielded the lowest selection bias of 16.10. Baldi allocation had relatively high selection bias of 101.50.

The percentiles of the realized selection bias values resulting from 10,000 simulated trial was utilized in deciding the shape of the individual desirability function. For example, the 90th percentile selection bias value was 119, and thus a selection bias value of 120 or greater will be given a score of 0. Selection bias of 65 was the 60th percentile of all realized values, and was given a score of 0.4; the value of 12 was the 20th percentile and contributed to the decision of giving a value of 10 a score of 0.8. Ultimately, the values (120, 100, 65, 20, 10, 0) were mapped to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1.0). Table 5.9 summarizes the distributions of the resulting individual desirability scores for selection bias.

	Individual Desirability Scores for Selection Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	1.000	1.000	1.000	1.000	1.000	1.000
PBD	0.319	0.408	0.421	0.420	0.432	0.478
ERADE.Neyman	0.000	0.015	0.029	0.033	0.046	0.152
ERADE.RSIHR	0.000	0.023	0.038	0.042	0.057	0.169
ERADE.Urn	0.000	0.000	0.008	0.016	0.026	0.152
ERADE.Baldi	0.000	0.000	0.000	0.000	0.000	0.000
DBCD.Neyman	0.352	0.586	0.611	0.629	0.672	0.824
DBCD.RSIHR	0.503	0.637	0.688	0.684	0.731	0.843
DBCD.Urn	0.217	0.497	0.527	0.521	0.551	0.740
DBCD.Baldi	0.000	0.128	0.186	0.174	0.225	0.360
SMLE.Neyman	0.416	0.745	0.814	0.795	0.861	0.947
SMLE.RSIHR	0.631	0.858	0.891	0.880	0.913	0.961
SMLE.Urn	0.205	0.543	0.583	0.592	0.641	0.866
SMLE.Baldi	0.000	0.182	0.207	0.198	0.221	0.266
EW1995.Neyman	0.408	0.636	0.705	0.698	0.760	0.876
EW1995.RSIHR	0.555	0.732	0.779	0.769	0.815	0.901
EW1995.Urn	0.000	0.514	0.548	0.541	0.574	0.748
EW1995.Baldi	0.000	0.152	0.192	0.181	0.216	0.303

Table 5.9: AZT Trial: summary statistics for individual desirability scores for selection bias under $H_1 : \beta_1 = 1.330$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

Since selection bias is always 0 for CRD, its individual desirability score for this component is always 1. The block size of PBD was 8, small enough such that it scored less well for this component, with an average score of 0.420. Baldi allocation targeted by ERADE always had selection bias greater than 120, so its individual desirability score is consistently 0 across all simulated trials. It is able to do better under other designs; for example, DBCD.Baldi had an average selection bias individual desirability score of 0.174. Neyman and RSIHR allocations did well with average scores ranging from 0.62 to 0.88, except when targeted by ERADE in which case selection bias was more probable, with scores less than 0.5

If the statisticians and clinicians in the planning phases of the trial believe that the risk of selection bias is minimal, it is recommended that the individual desirability function still be constructed with care to reflect preferences of penalization if selection bias were in fact a concern. The weight of the selection bias can be lowered accordingly, or even set to zero, so that a design's score on this component plays little or no role in the overall evaluation of a design.

Expected Number of Failures

In the context of this trial, minimizing the expected number of failures equates to minimizing the number of maternal-infant HIV transmissions. Due to the nature of the outcome, this ethical objective could be considered much more critical in this trial than in those with less extreme outcomes. While RSIHR allocation specifically aims to minimize the expected number of failures, we seek to evaluate the performance of the other designs. Table 5.10 shows the expected number of failures as deduced from 10,000 simulated trials.

	Expected Number of Failures		
	Min	Mean	Max
CRD	48.00	80.62	110.00
PBD	50.00	80.61	111.00
ERADE.Neyman	60.00	90.17	127.00
ERADE.RSIHR	50.00	78.40	110.00
ERADE.Urn	31.00	60.36	92.00
ERADE.Baldi	40.00	65.23	95.00
DBCD.Neyman	60.00	90.05	126.00
DBCD.RSIHR	50.00	78.43	110.00
DBCD.Urn	27.00	60.05	95.00
DBCD.Baldi	39.00	65.08	89.00
SMLE.Neyman	60.00	89.70	122.00
SMLE.RSIHR	46.00	78.46	110.00
SMLE.Urn	28.00	60.81	100.00
SMLE.Baldi	36.00	65.06	96.00
EW1995.Neyman	59.00	90.06	126.00
EW1995.RSIHR	49.00	78.46	112.00
EW1995.Urn	28.00	60.33	93.00
EW1995.Baldi	36.00	65.03	99.00

Table 5.10: AZT trial: expected number of failures under $H_1 : \beta_1 = 1.330$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

Recall that the objective of Neyman allocation is to maximize power for a fixed sample size, and that it is well-known that Neyman allocation can place less subjects in the stronger-performing arm when the probability of success is high in both arms of a trial. Neyman allocation's inability to place more subjects in the AZT experimental arm results in its inability to yield lower maternal-infant HIV transmission rates; Neyman allocations resulted in an average of 90 failures out of 477 subjects, regardless of the design implementing the allocation. Neyman allocation also had the highest worse-case scenarios of 127, 126, 122, and 126 when targeted by ERADE, DBCD, SMLE, and EW1995, respectively. We see then that in the case of the AZT trial, with high probabilities of success in both the AZT and control arms of 0.917 and 0.745, respectively, Neyman allocation performs worse with respect to expected number of failures than does non-RAR designs CRD and PBD. The other target allocations on average are able to reduce the number of expected failures compared to CRD and PBD. Whilst RSIHR's goal is to minimize the expected number of failures, we see that on average it is able to outperform Neyman allocation, but actually yields more failures than do Urn and Baldi allocations. Specifically, RSIHR allocation results in about 78 maternal-infant HIV transmissions, regardless of which RAR design targets it. This is certainly an improvement over Neyman's 90 expected failures. However, Baldi allocation results in an expected 65 maternal-infant HIV transmissions, and Urn allocation does extremely well with 60 expected maternal-infant HIV transmissions. This means that for this trial, Urn allocation would result in an expected 23% less maternal-infant HIV transmissions than RSIHR allocation. The best-case scenario witnessed in 10,000 simulated trials is also provided by Urn allocation, targeted by either SMLE or EW1995 designs, with 28 maternal-infant HIV transmissions out of 477 total subjects. Urn and Baldi share in worst-case scenarios, with maximum number of failures seen across 10,000 simulated trials ranging from 89 to 99.

In the construction of the individual desirability function for expected number of failures, the minimum and maximum number of failures across the 10,000 simulated trials was considered. Although the maximum number of failures witnessed was 127, the 90th percentile for expected number of failures was 93. Consequently, trials yielding 100 failures or more were given individual desirability scores for this component of 0. The 20th percentile for number of failures was 62, which was mapped to a score of 0.6. Continuing to evaluate the distribution of failures across the simulated trials, the values (100, 85, 72, 62, 50, 25) were given scores (0, 0.2, 0.4, 0.6, 0.8, 1.0). Table 5.11 summarizes the resulting individual desirability score distributions for each design.

	Individual Desirability Scores for Expected Number of Failures					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.000	0.187	0.262	0.274	0.354	0.816
PBD	0.000	0.187	0.262	0.273	0.354	0.800
ERADE.Neyman	0.000	0.053	0.133	0.141	0.200	0.633
ERADE.RSIHR	0.000	0.215	0.308	0.307	0.385	0.800
ERADE.Urn	0.107	0.520	0.633	0.617	0.717	0.952
ERADE.Baldi	0.067	0.440	0.540	0.533	0.633	0.880
DBCD.Neyman	0.000	0.053	0.133	0.142	0.215	0.633
DBCD.RSIHR	0.000	0.215	0.308	0.307	0.385	0.800
DBCD.Urn	0.067	0.520	0.633	0.622	0.733	0.984
DBCD.Baldi	0.147	0.440	0.540	0.536	0.633	0.888
SMLE.Neyman	0.000	0.053	0.133	0.148	0.215	0.633
SMLE.RSIHR	0.000	0.215	0.308	0.307	0.385	0.832
SMLE.Urn	0.000	0.500	0.617	0.608	0.717	0.976
SMLE.Baldi	0.053	0.440	0.540	0.536	0.633	0.912
EW1995.Neyman	0.000	0.053	0.133	0.142	0.215	0.650
EW1995.RSIHR	0.000	0.215	0.308	0.306	0.385	0.808
EW1995.Urn	0.093	0.520	0.633	0.617	0.717	0.976
EW1995.Baldi	0.013	0.440	0.540	0.537	0.633	0.912

Table 5.11: AZT Trial: summary statistics for individual desirability scores for expected number of failures under $H_1 : \beta_1 = 1.330$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

With trials resulting in 120 failures or more being considered absolutely unacceptable with an individual desirability score of 0, the majority of the designs assessed did have worst-case scenarios attaining this undesirable score. Specifically, all designs *except* ERADE.Urn, ERADE.Baldi, DBCD.Urn, DBCD.Baldi, SMLE.Baldi, EW1995.Urn, and EW1995.Baldi had at least one simulated trial out of 10,000 which resulted in 120 failures or more. While Baldi and urn allocations typically were able to have better worst-case scenarios, with non-zero minimum individual desirability scores, they also scored highest on average, with scores ranging between 0.53 and 0.63, compared with CRD and PBD's individual desirability score of 0.27. The designs with the worst best-case scenarios can be seen by their lowest maximum individual desirability scores: Neyman allocation even in their strongest trials with regards to minimizing failures had *maximum* scores of 0.633 to 0.650, which was close to urn allocation's *average* individual desirability scores. The large range and even difference in distributions of the individual desirability scores for expected number of failures prove it to be a strong differentiator between designs if this ethical objective is important to those designing the trial.

Bias

Table 5.12 displays bias under the null hypothesis across 10,000 simulated trials. Recall that the true difference in success probabilities under the null is 0, and $\boldsymbol{\beta} = (\beta_0, \beta_1) = (1.072121, 0)$.

	Bias $E(\hat{\beta}_1 - \beta_1)$		
	Min	Mean	Max
CRD	-0.740	0.002	0.729
PBD	-0.750	-0.000	0.868
ERADE.Neyman	-0.918	0.000	0.782
ERADE.RSIHR	-0.766	0.001	0.901
ERADE.Urn	-0.896	0.003	0.781
ERADE.Baldi	-1.054	-0.014	0.841
DBCD.Neyman	-0.823	0.004	0.797
DBCD.RSIHR	-0.790	-0.000	0.748
DBCD.Urn	-0.928	0.001	0.925
DBCD.Baldi	-1.162	-0.016	0.953
SMLE.Neyman	-0.820	0.002	0.848
SMLE.RSIHR	-0.825	0.002	0.750
SMLE.Urn	-0.917	0.002	0.784
SMLE.Baldi	-0.931	-0.009	0.913
EW1995.Neyman	-0.890	-0.001	0.816
EW1995.RSIHR	-0.754	0.002	0.861
EW1995.Urn	-0.816	0.002	0.818
EW1995.Baldi	-1.070	-0.005	0.857

Table 5.12: AZT trial: bias of the treatment effect estimate $(E(\hat{\beta}_1 - \beta_1))$ under $H_0 : \beta_1 = 0$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

The average bias under the null for most designs was close to 0. Specifically, PBD, ERADE.Neyman, and DBCD.RSIHR all had average bias of 0.000. CRD had an average bias of 0.002. Designs yielding the highest average bias were ERADE.Baldi and DBCD.Baldi, with a bias of -0.014 and -0.016, respectively. Note that CRD has the smallest range of bias, ranging from -0.740 to 0.729. This means CRDs worst underestimation of the treatment effect results in estimating the difference in success probabilities between AZT and placebo to be exp(1.072121 + 0 - 0.74)/(1+exp(1.072121 + 0 - 0.74)) - exp(1.072121)/(1+exp(1.072121)) = -0.163, and its worst overestimation of the treatment effect results in estimating the difference in success probabilities to be exp(1.072121 + 0 + 0.729)/(1+exp(1.072121 + 0 + 0.729)) - exp(1.072121)/(1+exp(1.072121)) = 0.113.

Amongst the four target allocations assessed, for a given design (e.g. DBCD), Baldi had the worst underestimation of β_1. Baldi targeted by ERADE, DBCD, SMLE, and EW1995 had a minimum bias of -1.054, -1.162, -0.931, and -1.070, respectively, across 10,000 simulated trials. For overestimation of β_1, DBCD.Baldi had the worst maximum bias of 0.953, followed closely by DBCD.Urn with a maximum bias of 0.925.

While the minimum bias of the treatment effect under the null hypothesis was -1.16 in 10,000 simulated trials, the 10th percentile bias value was -0.28. Similarly, the maximum bias realized was 0.95, but the 90th percentile bias value was +0.28. Given this, trials resulting in bias of ±0.5 are given an individual desirability score of 0 for the bias component. The value of 0 is given a score of 1. A symmetrical nominal-the-better individual desirability function is constructed based off the distributional characteristics of the realized bias values across 10,000 simulations. For example, the median bias was 0.0009 and the 45th and 55th percentiles were ±0.025 and assigned scores of 0.80. The final definition of the individual desirability function assigned values of (-0.5, -0.25, -0.12, -0.05, -0.025, 0, 0.025, 0.05, 0.12, 0.25, 0.5) to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1, 0.8, 0.6, 0.4, 0.2, 0). Table 5.13 displays the summary statistics of resulting individual desirability scores for bias.

	Individual Desirability Scores for Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.000	0.213	0.369	0.405	0.551	1.000
PBD	0.000	0.208	0.361	0.401	0.546	0.990
ERADE.Neyman	0.000	0.213	0.367	0.404	0.551	0.999
ERADE.RSIHR	0.000	0.211	0.371	0.402	0.546	0.991
ERADE.Urn	0.000	0.210	0.367	0.403	0.551	1.000
ERADE.Baldi	0.000	0.186	0.341	0.377	0.529	0.997
DBCD.Neyman	0.000	0.212	0.366	0.404	0.554	0.999
DBCD.RSIHR	0.000	0.211	0.369	0.404	0.552	0.999
DBCD.Urn	0.000	0.210	0.368	0.404	0.554	1.000
DBCD.Baldi	0.000	0.183	0.343	0.378	0.529	1.000
SMLE.Neyman	0.000	0.208	0.368	0.404	0.551	1.000
SMLE.RSIHR	0.000	0.211	0.368	0.407	0.555	1.000
SMLE.Urn	0.000	0.206	0.364	0.401	0.549	0.998
SMLE.Baldi	0.000	0.187	0.344	0.383	0.536	1.000
EW1995.Neyman	0.000	0.201	0.359	0.396	0.543	1.000
EW1995.RSIHR	0.000	0.204	0.365	0.400	0.547	1.000
EW1995.Urn	0.000	0.207	0.364	0.399	0.546	0.999
EW1995.Baldi	0.000	0.182	0.344	0.378	0.536	1.000

Table 5.13: AZT Trial: summary statistics for individual desirability scores for bias ($E(\hat{\beta}_1 - \beta_1)$) under $H_0 : \beta_1 = 0$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

Table 5.13 shows that each design's distribution of individual desirability scores for the bias component are quite similar; there is less variance in their summary statistics than was witnessed in the expected number of failures component. Each design has scores ranging between 0 and nearly 1. On average, most designs score around 0.40. CRD has an average score of 0.405; many designs are able to do almost as well. The designs that do more poorly on average are ones that target Baldi allocation, with an average score of approximately 0.38, and lower 25th, 50th, and 75th percentile values of 0.18, 0.34, and 0.53, respectively. However, the difference in the individual desirability scores is smaller than those seen in other components.

While bias under the null is important, relative bias under the alternative is probably of more interest to those designing the clinical trial, since they are inclined to believe a treatment effect does exist.

Relative Bias

Table 5.14 shows relative bias across 10,000 simulated trials for the evaluated designs.

	Relative Bias $E(\frac{\hat{\beta}_1-\beta_1}{\beta_1}) \times 100$		
	Min	Mean	Max
CRD	-77.70	1.13	111.91
PBD	-74.39	1.19	105.62
ERADE.Neyman	-72.52	6.47	1232.99
ERADE.RSIHR	-66.71	1.23	118.75
ERADE.Urn	-68.26	1.77	101.72
ERADE.Baldi	-70.66	1.05	99.46
DBCD.Neyman	-73.69	5.01	1232.09
DBCD.RSIHR	-75.78	1.29	106.87
DBCD.Urn	-77.62	1.77	124.31
DBCD.Baldi	-80.98	1.10	101.04
SMLE.Neyman	-72.25	4.13	1219.22
SMLE.RSIHR	-75.76	1.36	104.21
SMLE.Urn	-70.86	1.61	109.41
SMLE.Baldi	-73.71	0.68	94.78
EW1995.Neyman	-75.05	4.61	1229.39
EW1995.RSIHR	-67.88	1.40	86.27
EW1995.Urn	-70.88	1.79	101.13
EW1995.Baldi	-82.85	0.97	90.71

Table 5.14: AZT trial: bias of the treatment effect estimate under the alternative hypothesis ($E(\frac{\hat{\beta}_1-\beta_1}{\beta_1}) \times 100$) for various designs evaluating $\boldsymbol{p} = (0.917, 0.745), \boldsymbol{\beta} = (\beta_0, \beta_1) = (1.072121, 1.330146)$ and $n = 477$.

Since the minimum relative bias is still greater than -100, we see that none of the designs conclude that the placebo arm has a higher probability of success. SMLE targeting Baldi yields the lowest average relative bias of 0.68 across 10,000 simulated trials, even lower than the average relative bias of 1.13 yielded by CRD. The strong performance of SMLE targeting Baldi is not undone by a large variance of relative bias; its minimum and maximum relative biases are -73.71 and 94.78, respectively, both better than the minimum and maximum of CRD of -77.70 and 111.91, respectively. Baldi targeted by ERADE, DBCD, and EW1995 also have better average relative bias values than CRD. Note, however, that EW1995.Baldi yields worse minimum relative bias of -82.85, compared with -77.70 of CRD.

While Baldi is a strong performer with regards to relative bias, Neyman allocation yields higher relative biases. For example, Neyman allocation targeted by ERADE yielded an average relative bias of 6.47. This means that on average, ERADE.Neyman estimated $\hat{\beta}_1$ to be $0.0647 \times 1.330146 + 1.330146 = 1.416$, meaning that the estimated difference in probabilities of success in the two arms attributable to the treatment was 0.178 rather than the true difference in probabilities of 0.172. We can see that on average, then, these designs all perform quite well with regards to relative bias. Inspecting the maximum relative biases seen in 10,000 simulated trials, we see that most designs have a worst-case relative bias of about 100, which would mean that the design estimated $\hat{\beta}_1$ to be $1 \times 1.330146 + 1.330146 = 2.663$, resulting in an estimation of the difference in probabilities of success in the AZT versus placebo arms to be 0.977 - 0.745 = 0.232. ERADE.Neyman had a maximum relative bias of 1233, which means that it estimated $\hat{\beta}_1$ to be $12.33 \times 1.330146 + 1.330146 = 17.731$ (estimating guaranteed success in the AZT arm), resulting in an estimation of the difference in probabilities of success in the AZT versus placebo arms to be 0.999-0.745 = 0.255. Neyman's highest average relative bias and highest maximum relative bias across 10,000 simulated trials goes hand-in-hand with its goal to maximize power, since overestimation of the treatment effect would lead to a correct rejection of the null hypothesis and an increase in power.

The minimum relative bias observed across 10,000 simulated trials was nearly -83, and the 1st percentile was -0.45. This leads us to setting our minimum acceptable relative bias to -80. On the other hand, the maximum relative bias was nearly 1233, and the 99th percentile of relative bias was 55; the maximum acceptable relative bias is set to 100. Since under the alternative, there exists a positive treatment effect, positive values of relative bias ought to be penalized less than negative values. Observing the distributional characteristics of relative bias in the simulated trials, the values (-80, -50, -20, -11, -5, 0, 5, 20, 35, 60, 100) are mapped to individual desirability scores of (0, 0.2, 0.4, 0.6, 0.8, 1.0, 0.8, 0.6, 0.4, 0.2, 0). Table 5.15 summarizes the resulting individual desirability score distributions for relative bias. With our choice of relative bias value-to-score mapping, we see some differentiation between the performances of the designs, yet the scores do not vary as much as, say, those for expected number of failures. Other individual desirability functions were tested for sensitivity analyses but the similarity of the distributions for relative bias amongst the designs evaluated led to similar differences in scores for the other tested functions as well. Due to our definition of -80 and 100 as our minimally and maximally acceptable relative bias values, most designs yield scores ranging between 0 and 1. CRD has a mean individual desirability score for relative bias of 0.611, as does ERADE.RSIHR, ERADE.Urn, and DBCD.Urn. The only design that fairs better is DBCD.Baldi, with an average score of 0.612. Neyman allocation does not do as well, with a slightly lower score ranging from 0.58 to

	Individual Desirability Scores for Relative Bias					
	Min	25th Percentile	Median	Mean	75th Percentile	Max
CRD	0.000	0.434	0.617	0.611	0.771	1.000
PBD	0.000	0.430	0.615	0.608	0.769	0.998
ERADE.Neyman	0.000	0.394	0.590	0.583	0.753	1.000
ERADE.RSIHR	0.000	0.433	0.623	0.611	0.770	1.000
ERADE.Urn	0.000	0.426	0.611	0.605	0.764	1.000
ERADE.Baldi	0.003	0.436	0.613	0.611	0.772	0.999
DBCD.Neyman	0.000	0.396	0.594	0.588	0.761	1.000
DBCD.RSIHR	0.000	0.438	0.622	0.611	0.769	1.000
DBCD.Urn	0.000	0.420	0.611	0.604	0.765	1.000
DBCD.Baldi	0.000	0.439	0.621	0.612	0.769	1.000
SMLE.Neyman	0.000	0.398	0.595	0.588	0.757	1.000
SMLE.RSIHR	0.000	0.435	0.620	0.610	0.767	1.000
SMLE.Urn	0.000	0.425	0.612	0.605	0.764	1.000
SMLE.Baldi	0.026	0.440	0.625	0.614	0.771	1.000
EW1995.Neyman	0.000	0.397	0.591	0.586	0.756	1.000
EW1995.RSIHR	0.069	0.444	0.628	0.615	0.769	1.000
EW1995.Urn	0.000	0.424	0.614	0.605	0.762	1.000
EW1995.Baldi	0.000	0.446	0.628	0.615	0.771	1.000

Table 5.15: AZT Trial: summary statistics for individual desirability scores for relative bias $E(\frac{\hat{\beta_1}-\beta_1}{\beta_1}) \times 100$ under $H_1 : \beta_1 = 1.330$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

0.59. Neyman's poor performance is also apparent when looking at 25th percentile values: SMLE.Neyman's 25th percentile score is 0.397, compared with ERADE.Baldi, which has the highest 25th percentile score of 0.446. 75th percentile values also show little variance, ranging from 0.753 by ERADE.Neyman to 0.772 by ERADE.Baldi. EW1995.RSIHR's minimum score of 0.069 shows less extreme biases than most of the other designs with a minimum score of 0. However, the other designs targeting RSIHR - ERADE, DBCD, SMLE - were not able to avoid scores of 0 due to their maximum relative bias values being greater than 100.

Type I Error and Power

While RAR provide value in their ability to target allocations that fulfill objectives such as maximizing power or minimizing failures, the FDA has emphasized that designs must control the Type I error. The proportion of trials amongst the 10,000 simulated that resulted in a failure to reject the null hypothesis is the Type I error, and the proportion of trials that resulted in the correct rejection of the null hypothesis is the power. Table 5.16 displays the Type I error and power.

The ability of all designs evaluated to achieve power greater than 90% is not surprising, since with $p_E = 0.917$ and $p_C = 0.745$, only 189 subjects are needed to achieve 90% power at the $\alpha = 0.05$ level. The overwhelming evidence of the treatment effect is what resulted in the trial's early termination after its first interim look. While power is strong for designs evaluated, we can see Type I error varies from 0.0486 of SMLE.Urn, to 0.539 of EW1995.Neyman. Note that EW1995 consistently has higher Type I error greater than 0.053. For example, EW1995 targeting urn allocation had Type I error of 0.0535, even though

	Type I Error	Power
CRD	0.0493	0.9995
PBD	0.0502	0.9995
ERADE.Neyman	0.0479	0.9973
ERADE.RSIHR	0.0506	0.9995
ERADE.Urn	0.0504	0.9995
ERADE.Baldi	0.0513	0.9985
DBCD.Neyman	0.0500	0.9979
DBCD.RSIHR	0.0518	0.9994
DBCD.Urn	0.0482	0.9983
DBCD.Baldi	0.0485	0.9978
SMLE.Neyman	0.0522	0.9987
SMLE.RSIHR	0.0525	0.9992
SMLE.Urn	0.0486	0.9981
SMLE.Baldi	0.0502	0.9976
EW1995.Neyman	0.0539	0.9983
EW1995.RSIHR	0.0536	0.9995
EW1995.Urn	0.0535	0.9987
EW1995.Baldi	0.0535	0.9978

Table 5.16: AZT trial: Type I error and power for various designs evaluating $\boldsymbol{p} = (0.917, 0.745), \boldsymbol{\beta} = (\beta_0, \beta_1) = (1.072121, 1.330146)$ and $n = 477$.

SMLE targeting the same urn allocation was able to keep Type I error at 0.0486. The designs that were able to control their Type I error at or below 5% are CRD, ERADE.Neyman, DBCD.Neyman, DBCD.Urn, DBCD.Baldi, and SMLE.Urn.

The individual desirability function definitions for Type I error and power are easier to define, since most statisticians have a clear conception of what they consider acceptable or not. This trial wants to control Type I error at 5%, but Table 5.16 reveals that some designs yield lower Type I errors. Given this, we will assign a Type I error of 0.05 the individual desirability score of 0.8, highly acceptable, and the Type I error of 0.045 or lower the individual desirability score of 1. As 0.0539 is the largest Type I error, provided by the EW1995.Neyman, the Type I error of 0.054 and higher is given the score of 0. Given this, we decide to penalize deviations from 0.05 to 0.054 equally: the Type I error values of (0.054, 0.053, 0.052, 0.051, 0.05, 0.045) are given scores (0, 0.2, 0.4, 0.6, 0.8, 1).

The evaluation of power can be done in two ways. Since each of the designs evaluated resulted in power greater than 0.9, we can automatically assign them all individual desirability scores of 1 for the power component. The other method is to relate power with sample size and score the designs with higher power with a higher score since this could be a proxy function for sample size needed. We will evaluate sample size needed separately in this case. A desirability score of 0.9 is desired, so a design yielding power of 0.9 or higher is given an individual desirability score of 1 for the component of power. A desirability score below 0.8 is considered absolutely unacceptable, so the value of 0.79 or less is given a score of 0. More specifically, the values (0.79, 0.80, 0.83, 0.85, 0.88, 0.9) to scores (0, 0.2, 0.4, 0.6, 0.8, 1). Table 5.17 displays the resulting individual desirability scores for Type I error and power for each of the designs considered.

	Individual Desirability Scores	
	Type I Error	Power
CRD	0.828	1.000
PBD	0.760	1.000
ERADE.Neyman	0.884	1.000
ERADE.RSIHR	0.680	1.000
ERADE.Urn	0.720	1.000
ERADE.Baldi	0.540	1.000
DBCD.Neyman	0.800	1.000
DBCD.RSIHR	0.440	1.000
DBCD.Urn	0.872	1.000
DBCD.Baldi	0.860	1.000
SMLE.Neyman	0.360	1.000
SMLE.RSIHR	0.300	1.000
SMLE.Urn	0.856	1.000
SMLE.Baldi	0.760	1.000
EW1995.Neyman	0.020	1.000
EW1995.RSIHR	0.080	1.000
EW1995.Urn	0.100	1.000
EW1995.Baldi	0.100	1.000

Table 5.17: AZT Trial: individual desirability scores for Type I error and power for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n = 477$.

Note that since only 180 subjects are needed to attain 90% level at the alpha = 0.05 level, each of the simulated trials has high power well over 0.90, resulting in each design having the highest individual desirability score of 1 for the power component. On the other hand, individual desirability scores for Type I error vary from 0.02 from EW1995.Neyman to 0.884 from ERADE.Neyman. EW1995.Neyman did not perform well with respect to Type I error. Conversely, DBCD.Neyman and ERADE.Neyman have Type I error individual desirability scores of 0.800 and 0.884, respectively. Urn allocation targeted by DBCD does next best, with a score of 0.872. SMLE.Urn also did well with a score of 0.85. Baldi allocation targeted by DBCD scores similarly with a score of 0.860. RSIHR allocation performs moderately with ERADE and a score of 0.680, and poorly when targeted by DBCD, SMLE, and RSIHR, with scores of 0.440, 0.300, and 0.080, respectively.

Overall Desirability Score

In order to calculate the overall desirability score, the relative importance of the 11 assessed characteristics is considered. The weights for treatment group size imbalance, expected number of failures, covariate imbalance C1, covariate imbalance C2, covariate imbalance C3, selection bias, accidental bias factor estimate, bias under the null hypothesis, relative bias under the alternative hypothesis, Type I error, power, and sample

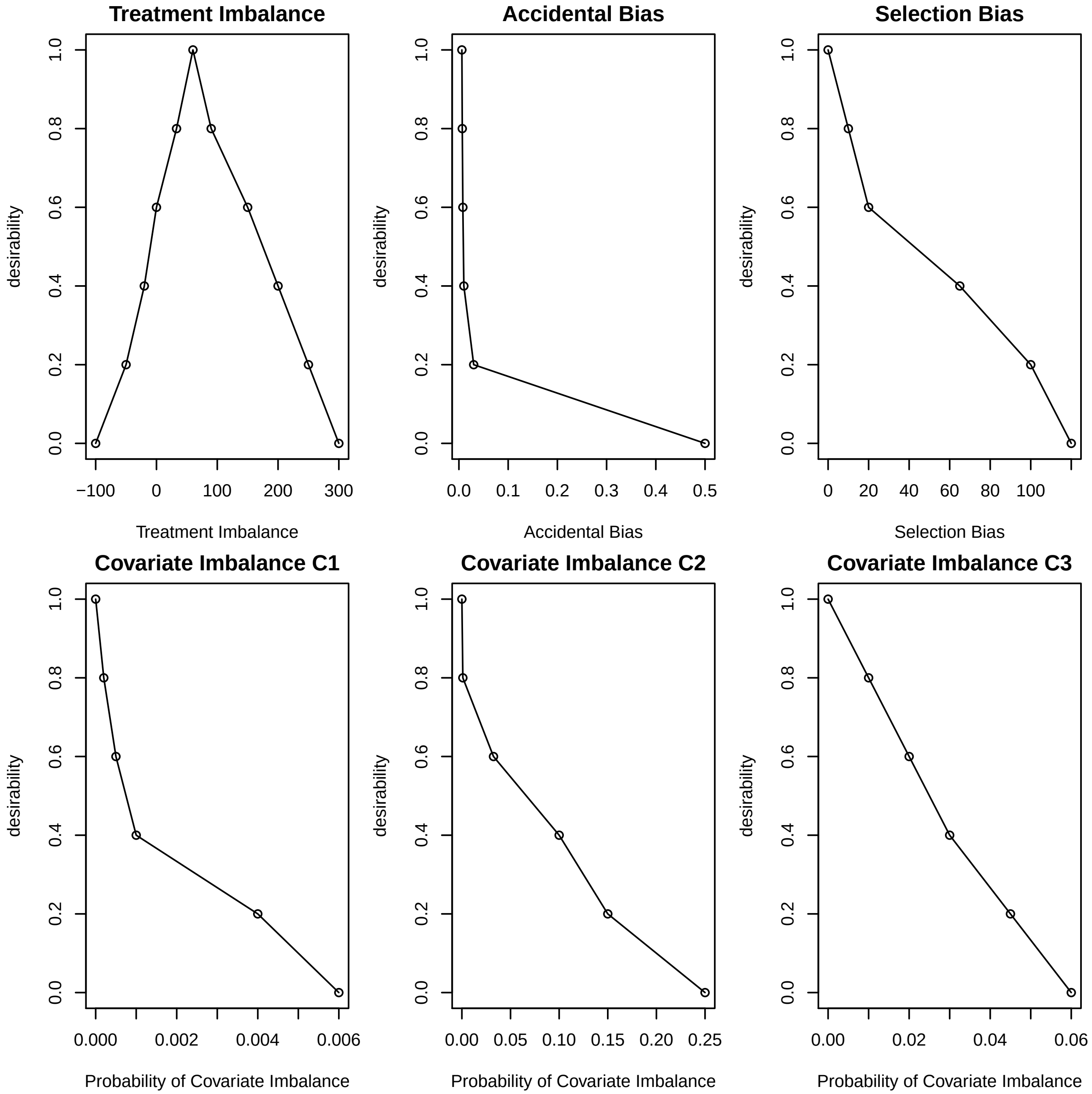


Figure 5.1: AZT trial: individual desirability functions.

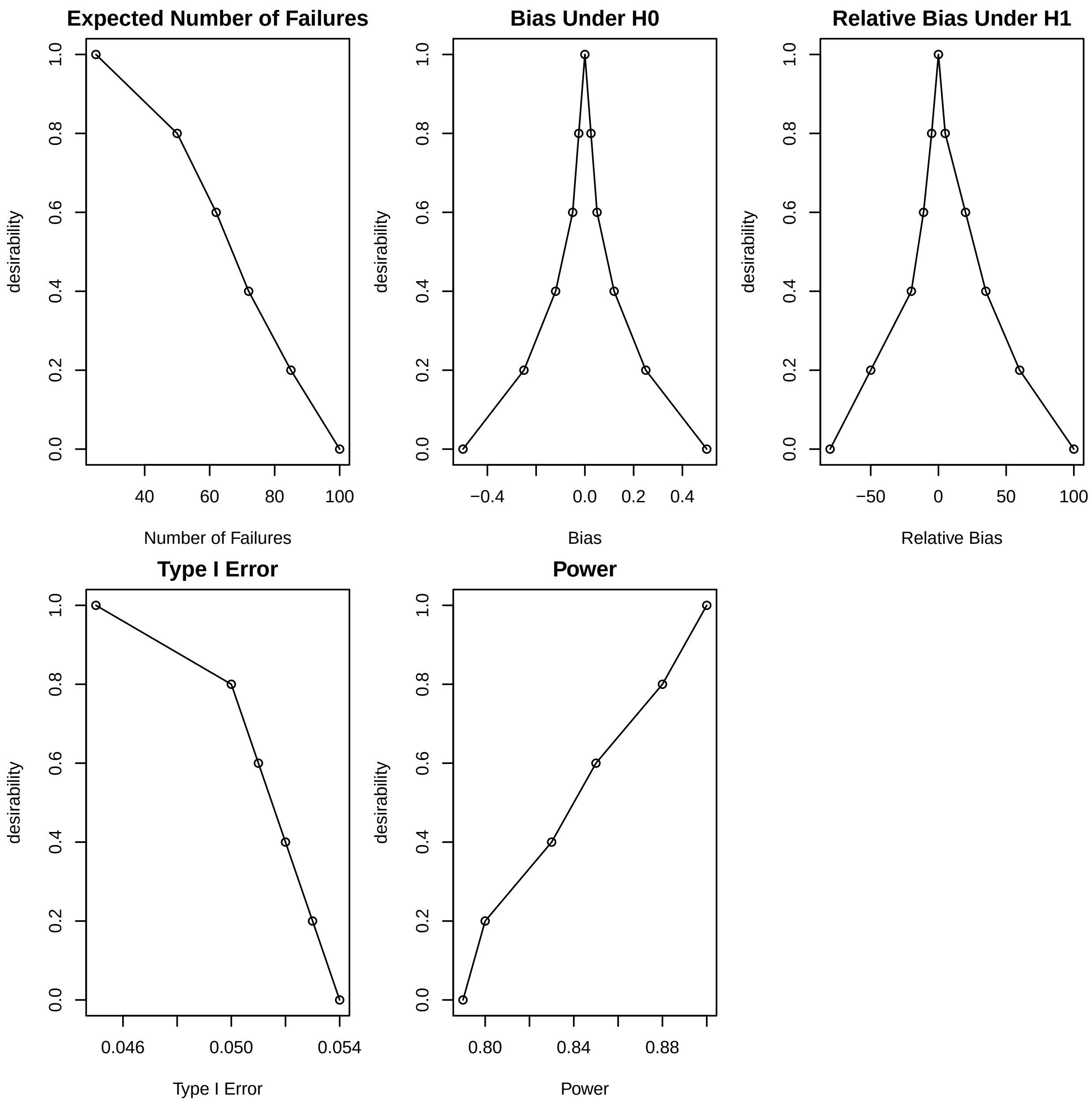


Figure 5.2: AZT trial: individual desirability functions, continued.

size needed to obtain 90% power are denoted in the vector

$$\boldsymbol{w} = (w_{\text{imbal}}, w_{\text{fails}}, w_{\text{c1}}, w_{\text{c2}}, w_{\text{c3}}, w_{\text{sb}}, w_{\text{accbias}}, w_{\text{bias}}, w_{\text{relbias}}, w_{\text{alpha}}, w_{\text{power}}, w_{\text{n}}).$$

In the first weight setting considered, the first decision is to place no weight on selection bias, because the trial is double-blinded and there is no reason to suspect the integrity of the blinding would be violated. The second decision is to place little weight on a) treatment group size imbalance and on b) bias under the null hypothesis. The driver for a decision like this would be when there is strong evidence in pre-clinical and early phase studies for efficacy of experimental drug AZT, so a) placing more subjects in the AZT arm might be preferred for ethical reasons, and b) there is strong evidence that the trial will show preference to the alternative hypothesis, reducing the importance of bias under the null. Next, the primary endpoint of maternal-infant transmission of HIV Type I, determined by the 18-month followup of the infant is considered to be highly undesirable, so we want to prioritize minimizing the expected number of failures and place heavy weight on this component. Since a time trend in confounders may play a role, some weight is given to covariate imbalance C2. As discussed in Section 4.1.3, covariate type C3 in healthcare settings often represents some association with pollutants or other exposure that follows an autocorrelated lag model. Although there is no certainty whether this exists, we can hedge against imbalance of a covariate of this type by placing some weight on this component. More importantly, heavier weight is assigned to balancing covariate type C1, an often present normally distributed random variable. Since this trial's findings will be submitted to regulatory agencies for approval, control of Type I error is essential, and is given as much weight as minimizing the number of failures. Power is also important, but as many more subjects are being enrolled than necessary for 90% power, only little weight is assigned to this component.

To summarize, the weight settings are $\boldsymbol{w_1} = (w_{\text{imbal}} = 0.5, w_{\text{fails}} = 3, w_{\text{c1}} = 2, w_{\text{c2}} = 1, w_{\text{c3}} = 1, w_{\text{sb}} = 0, w_{\text{accbias}} = 3, w_{\text{bias}} = 0.5, w_{\text{relbias}} = 2, w_{\text{alpha}} = 3, w_{\text{power}} = 1)$.

With the individual desirability functions defined in this section and these weights, the highest scoring design was Permuted Block Design (PBD). Recall that the block size used throughout this work for PBD is 8. Tables 5.18 to 5.20 summarize the mean of the individual desirability scores for each component and that design component's weight, and display the resulting overall desirability score for each design.

	CRD	PBD	ERADE.Neyman	ERADE.RSIHR	ERADE.Urn	ERADE.Baldi	weight
Treatment Group Size Imbalance	0.573	0.597	0.027	0.749	0.261	0.485	0.029
Expected No. of Failures	0.274	0.273	0.141	0.307	0.617	0.533	0.176
Covariate Imbalance C1	1.000	1.000	0.327	0.560	0.120	0.360	0.118
Covariate Imbalance C2	0.608	0.799	0.398	0.793	0.000	0.716	0.059
Covariate Imbalance C3	0.560	0.682	0.508	0.772	0.000	0.534	0.059
Selection Bias	1.000	0.420	0.033	0.042	0.016	0.000	0.000
Accidental Bias	0.855	0.794	0.389	0.588	0.236	0.589	0.176
Bias	0.405	0.401	0.404	0.402	0.403	0.377	0.029
Relative Bias	0.611	0.608	0.583	0.611	0.605	0.611	0.118
Type I Error	0.828	0.760	0.884	0.680	0.720	0.540	0.176
Power	1.000	1.000	1.000	1.000	1.000	1.000	0.059
Overall Desirability D (mean)	0.620	0.621	0.000	0.559	0.000	0.540	
Prob(Overall Desirability D = 0)	0.028	0.028	0.648	0.022	1.000	0.037	

Table 5.18: AZT trial: mean individual desirability scores for 11 considered design characteristics, mean overall desirability score D, and Probability(D=0), for CRD, PBD, ERADE.Neyman, ERADE.RSIHR, ERADE.Urn, and ERADE.Baldi designs ($n = 477$).

	DBCD.Neyman	DBCD.RSIHR	DBCD.Urn	DBCD.Baldi	SMLE.Neyman	SMLE.RSIHR	weight
Treatment Group Size Imbalance	0.033	0.752	0.249	0.478	0.057	0.742	0.029
Expected No. of Failures	0.142	0.307	0.622	0.536	0.148	0.307	0.176
Covariate Imbalance C1	0.387	0.440	0.160	0.320	0.380	0.600	0.118
Covariate Imbalance C2	0.386	0.731	0.000	0.617	0.453	0.595	0.059
Covariate Imbalance C3	0.472	0.620	0.105	0.341	0.506	0.632	0.059
Selection Bias	0.629	0.684	0.521	0.174	0.795	0.880	0.000
Accidental Bias	0.398	0.864	0.260	0.626	0.222	0.861	0.176
Bias	0.404	0.404	0.404	0.378	0.404	0.407	0.029
Relative Bias	0.588	0.611	0.604	0.612	0.588	0.610	0.118
Type I Error	0.800	0.440	0.872	0.860	0.360	0.300	0.176
Power	1.000	1.000	1.000	1.000	1.000	1.000	0.059
Overall Desirability D (mean)	0.000	0.529	0.000	0.564	0.000	0.506	
Prob(Overall Desirability D = 0)	0.632	0.022	1.000	0.037	0.580	0.023	

Table 5.19: AZT trial: mean individual desirability scores for 11 considered design characteristics, mean overall desirability score D, and Probability(D=0), for DBCD.Neyman, DBCD.RSIHR, DBCD.Urn, DBCD.Baldi, SMLE.Neyman, and SMLE.RSIHR designs ($n = 477$)

	SMLE.Urn	SMLE.Baldi	EW1995.Neyman	EW1995.RSIHR	EW1995.Urn	EW1995.Baldi	weight
Treatment Group Size Imbalance	0.286	0.476	0.034	0.753	0.264	0.471	0.029
Expected No. of Failures	0.608	0.536	0.142	0.306	0.617	0.537	0.176
Covariate Imbalance C1	0.050	0.340	0.387	0.560	0.110	0.307	0.118
Covariate Imbalance C2	0.000	0.550	0.368	0.710	0.000	0.567	0.059
Covariate Imbalance C3	0.132	0.376	0.528	0.556	0.085	0.361	0.059
Selection Bias	0.592	0.198	0.698	0.769	0.541	0.181	0.000
Accidental Bias	0.308	0.602	0.326	0.865	0.287	0.632	0.176
Bias	0.401	0.383	0.396	0.400	0.399	0.378	0.029
Relative Bias	0.605	0.614	0.586	0.615	0.605	0.615	0.118
Type I Error	0.856	0.760	0.020	0.080	0.100	0.100	0.176
Power	1.000	1.000	1.000	1.000	1.000	1.000	0.059
Overall Desirability D (mean)	0.000	0.552	0.000	0.400	0.000	0.384	
Prob(Overall Desirability D = 0)	1.000	0.034	0.636	0.027	1.000	0.038	

Table 5.20: AZT trial: mean individual desirability scores for 11 considered design characteristics, mean overall desirability score D, and Probability(D=0), for SMLE.Urn, SMLE.Baldi, EW1995.Neyman, EW1995.RSIHR, EW1995.Urn, and EW1995.Baldi designs ($n = 477$).

In a sensitivity analysis, what sort of weight preference would lead us to choose a response-adaptive randomization? Tweaking two of the weights such that slightly more weight was placed on imbalance, and no weight was placed on covariate imbalance C1, such that $\boldsymbol{w_2} = (w_{\text{imbal}} = 1, w_{\text{fails}} = 3, w_{\text{c1}} = 0, w_{\text{c2}} = 1, w_{\text{c3}} = 1, w_{\text{sb}} = 0, w_{\text{accbias}} = 3, w_{\text{bias}} = 0.5, w_{\text{relbias}} = 2, w_{\text{alpha}} = 3, w_{\text{power}} = 1)$, Baldi allocation targeted by DBCD has the highest overall desirability score of 0.604.

On the other hand, if the trial is not blinded throughout the entire study's randomization period, or if the integrity of the blinding is questioned, more weight may be placed on selection bias. If we also emphasize importance of balancing covariate C1, place no weights on covariate imbalances C2 and C3, decrease weight placed on accidental bias factor estimate, and increase weight on power, such that $\boldsymbol{w_3} = (w_{\text{imbal}} = 1, w_{\text{fails}} = 3, w_{\text{c1}} = 2, w_{\text{c2}} = 0, w_{\text{c3}} = 0, w_{\text{sb}} = 0.5, w_{\text{accbias}} = 3, w_{\text{bias}} = 0.5, w_{\text{relbias}} = 3, w_{\text{alpha}} = 3, w_{\text{power}} = 2)$, Complete Randomized Design (CRD) has the highest overall desirability score of 0.642.

Table 5.21 displays only, for brevity, the overall desirability score and the probability of overall desirability being 0.

	$\boldsymbol{w_2}$		$\boldsymbol{w_3}$	
	D	Prob(D = 0)	D	Prob(D = 0)
CRD	0.580	0.028	0.642	0.028
PBD	0.583	0.028	0.599	0.028
ERADE.Neyman	0.000	0.648	0.000	0.663
ERADE.RSIHR	0.564	0.022	0.478	0.026
ERADE.Urn	0.000	1.000	0.331	0.421
ERADE.Baldi	0.567	0.037	0.000	1.000
DBCD.Neyman	0.000	0.632	0.000	0.632
DBCD.RSIHR	0.548	0.022	0.534	0.022
DBCD.Urn	0.000	1.000	0.481	0.107
DBCD.Baldi	0.604	0.037	0.551	0.050
SMLE.Neyman	0.000	0.580	0.000	0.580
SMLE.RSIHR	0.501	0.023	0.525	0.023
SMLE.Urn	0.000	1.000	0.429	0.122
SMLE.Baldi	0.585	0.034	0.547	0.034
EW1995.Neyman	0.000	0.636	0.000	0.636
EW1995.RSIHR	0.390	0.027	0.413	0.027
EW1995.Urn	0.000	1.000	0.321	0.131
EW1995.Baldi	0.398	0.038	0.381	0.040

Table 5.21: AZT trial: mean overall desirability score D, and Probability(D = 0), for 18 considered designs ($n = 477$).

5.3 Design Selection with Sample Size Reduction

While the previous analysis of the AZT trial utilized the same sample size as the trial in practice, we observed that the study was highly powered, leading to its early termination in favor of investigative treatment approval after the first interim analysis. We concluded that PBD, DBCD.Baldi, and CRD were the most desirable given the three stated preference settings of the 11 design characteristics of interest.

However, given that these designs actually require different sample sizes to achieve 90% power, which design performs best with respect to these same design characteristics of interest, after reaching a sample size sufficient to obtain 90% power? To answer this question, simulation was used to obtain the sample size needed for each design. Other than sample size, the same design parameters as presented in the previous section are utilized. Using the necessary sample size only, the analysis is repeated. The results are detailed in this section.

Treatment Group Characteristics

Table 5.22 summarizes treatment group characteristics. Note that, in general, non-RAR designs require smaller sample size to obtain 90% power. This aligns with our expectations, since the variability in the changing target allocation of RAR designs reduces power (see Section 1.2.3). ERADE is the most efficient of the RAR designs discussed, and thus has smaller sample sizes than the other RAR designs. The sample size needed to obtain 90% power ranges from 190 for ERADE.Neyman and ERADE.RSIHR designs, to 226 for DBCD.Baldi design. The proportion of subjects placed in the experimental AZT arm ranges from 0.38 (Neyman allocation targeted by ERADE, DBCD, and EW1995) to 0.75 (Urn allocation targeted by DBCD). These allocations are consistent with those shown in Table 5.1 with the full sample size of $n = 477$.

	n Needed for 90% power	Patients in E (mean)	Patients in E (sd)	Proportion in E
CRD	195	97.70	6.98	0.50
PBD	195	97.50	0.80	0.50
ERADE.Neyman	190	71.94	10.49	0.38
ERADE.RSIHR	190	99.92	1.75	0.53
ERADE.Urn	210	154.87	13.95	0.74
ERADE.Baldi	218	150.34	3.77	0.69
DBCD.Neyman	193	73.64	10.86	0.38
DBCD.RSIHR	193	101.49	3.61	0.53
DBCD.Urn	215	160.41	15.06	0.75
DBCD.Baldi	226	156.63	4.94	0.69
SMLE.Neyman	195	77.02	11.90	0.39
SMLE.RSIHR	195	102.54	7.14	0.53
SMLE.Urn	215	156.74	17.59	0.73
SMLE.Baldi	225	155.49	7.91	0.69
EW1995.Neyman	195	74.04	11.55	0.38
EW1995.RSIHR	196	103.11	4.61	0.53
EW1995.Urn	220	162.39	16.67	0.74
EW1995.Baldi	220	152.91	6.42	0.70

Table 5.22: AZT trial reassessment using sample size needed for 90% power: Treatment group characteristics under $H_1 : \beta_1 = 1.330$ over 10,000 simulated trials evaluating $\boldsymbol{p} = (0.917, 0.745), n =$ n needed for 90% power.

	Treatment Group Size Imbalance $n_E - n_C$			
	Min	Median	Mean	Max
CRD	-51	1	0	57
PBD	-3	-1	-0	3
ERADE.Neyman	-116	-42	-46	22
ERADE.RSIHR	-2	10	10	30
ERADE.Urn	-10	100	100	192
ERADE.Baldi	68	80	83	140
DBCD.Neyman	-121	-43	-46	15
DBCD.RSIHR	-19	9	10	35
DBCD.Urn	-25	107	106	195
DBCD.Baldi	48	86	87	142
SMLE.Neyman	-121	-39	-41	53
SMLE.RSIHR	-47	11	10	67
SMLE.Urn	-65	101	98	191
SMLE.Baldi	23	85	86	159
EW1995.Neyman	-123	-43	-47	31
EW1995.RSIHR	-28	10	10	46
EW1995.Urn	-40	106	105	196
EW1995.Baldi	40	86	86	148

Table 5.23: AZT trial reassessment with reduced sample size: summary statistics for treatment group size imbalance, $n_E - n_C$, under $H_1 : \beta_1 = 1.330$ over 10,000 simulated trials evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n =$ n needed for 90% power.

Table 5.23 displays summary statistics for treatment group size imbalance $n_E - n_C$. The average treatment group size imbalance ranges from -47 as seen in the trial utilizing EW1995.Neyman, to +106, as seen in the trial utilizing DBCD.Urn. PBD has the smallest range in treatment group size imbalance of -3 to +3. The largest range is seen in EW1995.Urn, which places as many as 40 more subjects in the control arm to 196 more subjects in the experimental AZT arm.

Accidental Bias

Table 5.24 shows the multiplicative increases of the accidental relative bias factor summary statistics when reducing the sample size. For example, the average accidental bias factor estimate for CRD increased more than five-fold from 0.007 to 0.034 (five-fold calculated pre-rounding). The multiplicative increases when using sample size needed to obtain 90% power relative to the $n = 477$ of the original trial are shown in the parentheses. The multiplicative increases range from 3.8 when using SMLE.Baldi design, to 27.7 when using ERADE.Urn design. This large increase of 27.7 in maximum accidental bias factor estimate is due to the large treatment group size imbalance. While the proportion of subjects in experimental AZT arm is about 0.74 in both the previous and current analyses, the smaller sample size in the current analysis implies a larger treatment group size imbalance. Specifically, the largest witnessed proportion of imbalance in the reduced sample size analysis is $192/210 = 0.91$, larger than the proportion of imbalance in the original sample size analysis of $375/477 = 0.79$. (See Section 4.1.2.)

	Accidental Bias Factor Estimates		
	Min (Multiplicative Increase)	Mean (Multiplicative Increase)	Max (Multiplicative Increase)
CRD	0.034 (5.1)	0.034 (5.2)	0.040 (5.7)
PBD	0.036 (5.2)	0.036 (5.2)	0.036 (5.2)
ERADE.Neyman	0.072 (5.6)	0.086 (5.9)	0.183 (3.9)
ERADE.RSIHR	0.048 (5.8)	0.048 (5.8)	0.051 (6.0)
ERADE.Urn	0.094 (4.8)	0.183 (5.3)	3.487 (27.7)
ERADE.Baldi	0.033 (4.3)	0.037 (4.4)	0.078 (6.4)
DBCD.Neyman	0.071 (4.7)	0.084 (6.0)	0.194 (4.6)
DBCD.RSIHR	0.035 (5.5)	0.036 (5.5)	0.038 (5.6)
DBCD.Urn	0.108 (4.6)	0.226 (5.1)	3.424 (14.8)
DBCD.Baldi	0.027 (3.8)	0.034 (4.2)	0.066 (5.5)
SMLE.Neyman	0.082 (4.6)	0.094 (4.7)	0.217 (4.2)
SMLE.RSIHR	0.035 (5.3)	0.035 (5.4)	0.045 (6.1)
SMLE.Urn	0.143 (4.1)	0.285 (4.3)	3.218 (7.6)
SMLE.Baldi	0.025 (3.8)	0.035 (4.1)	0.099 (6.7)
EW1995.Neyman	0.079 (5.5)	0.094 (5.7)	0.217 (4.2)
EW1995.RSIHR	0.034 (5.2)	0.034 (5.2)	0.038 (5.5)
EW1995.Urn	0.121 (4.2)	0.248 (4.4)	2.850 (6.5)
EW1995.Baldi	0.027 (4.1)	0.036 (4.5)	0.085 (7.0)

Table 5.24: AZT trial reassessment with reduced sample size: accidental bias factor estimate under $H_1 : \beta_1 = 1.330$ over 10,000 simulated trials evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n =$ n needed for 90% power; multiplicative increases relative to accidental bias factor estimate yielded from $n = 477$ are shown in parentheses.

In the original analysis, we focused on the worst-case of an accidental bias factor estimate of EW1995.Urn (0.436 when $n = 477$), which resulted in underestimating the difference in probability of success by 0.031. With reduced sample sizes, EW1995.Urn still yielded the worst-case accidental bias factor estimate, yet it was much higher at 3.487. This means that for a covariate which results in a 5% decrease in probability of success with a one unit increase of the covariate ($\beta_{omitted} = -0.5275$), the bias on the treatment effect would be $\pm\sqrt{(-0.5275)^2 \times 3.487} = \pm 0.985$. Recall that the true value of $\beta_1 = 1.330146$, so the estimated difference in probability of success attributable to experimental treatment AZT would be $\frac{exp(1.072121+1.330146\pm0.985)}{1+exp(1.072121+1.330146\pm0.985)} - 0.745 = (0.805 - 0.745) or (0.967 - 0.745) = 0.060 or 0.222$, rather than the true difference in probability of success of 0.172.

The ranking of the designs using solely accidental bias factor estimate remained similar. For example, CRD, PBD, DBCD.RSIHR, SMLE.RSIHR, and EW1995.RSIHR once again had the lowest average accidental bias factor estimates. On the other hand, Urn allocation targeted by any of the RAR designs evaluated were the worst-performing designs with regards to this component, being the only designs yielding accidental bias factor estimates greater than 2.

Covariate Imbalance

Recall the three types of covariates discussed in Section 4.1.3: C1 is a standard normal variable, C2 represents a covariate that changes linearly over time, and C3 represents an autocorrelated variable. Table 5.25 display the probabilities of covariate imbalance exceeding 0.3 for these three covariates under the null and alternative hypotheses.

	Under H_0			**Under H_1**		
	C1	C2	C3	C1	C2	C3
CRD	0.037	0.173	0.135	0.000	0.173	0.135
PBD	0.033	0.039	0.132	0.000	0.039	0.132
ERADE.Neyman	0.043	0.116	0.119	0.040	0.262	0.149
ERADE.RSIHR	0.039	0.055	0.102	0.032	0.053	0.101
ERADE.Urn	0.031	0.283	0.144	0.056	0.492	0.222
ERADE.Baldi	0.042	0.077	0.137	0.036	0.097	0.127
DBCD.Neyman	0.037	0.151	0.142	0.039	0.253	0.158
DBCD.RSIHR	0.038	0.105	0.143	0.030	0.101	0.133
DBCD.Urn	0.029	0.303	0.136	0.060	0.456	0.195
DBCD.Baldi	0.040	0.114	0.149	0.037	0.131	0.143
SMLE.Neyman	0.037	0.177	0.140	0.034	0.237	0.153
SMLE.RSIHR	0.037	0.164	0.140	0.030	0.164	0.136
SMLE.Urn	0.030	0.245	0.125	0.053	0.414	0.187
SMLE.Baldi	0.039	0.210	0.146	0.036	0.180	0.140
EW1995.Neyman	0.035	0.163	0.135	0.041	0.265	0.153
EW1995.RSIHR	0.034	0.126	0.136	0.032	0.124	0.132
EW1995.Urn	0.026	0.292	0.123	0.055	0.427	0.187
EW1995.Baldi	0.043	0.182	0.153	0.036	0.158	0.157

Table 5.25: AZT trial reassessment with reduced sample size: probability of covariate imbalance under $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 = 1.330$ over 10,000 simulated trials, as defined by $|\overline{C}_E - \overline{C}_C| > 0.3, C \in \{C1, C2, C3\}$ for various designs evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n =$ n needed for 90% power.

The increase in probability of covariate imbalance with reduced sample size can be seen by comparing Tables 5.6 and 5.25. The largest increase in probability is seen in C1 under the null hypothesis, with a multiplicative increase of nearly 50 times, from 0.002 to 0.03. While this is a large increase, the probability of covariate imbalance for C1 under the null hypothesis is still small, with all designs yielding a probability less than 5%. Indeed, the average multiplicative increases under the null hypothesis were 39, 9, and 6 for C1, C2, and C3 covariates, respectively. Under the alternative hypothesis, SMLE.RSIHR saw the largest increase in probability of imbalance for C1, with a multiplicative increase greater than 59 from 0.0005 to 0.030. Again, although the increase is large, the overall probability of imbalance for C1 under the null is small, with the largest probability of covariate imbalance being 0.056 from ERADE.Urn. The average multiplicative increases under the alternative hypothesis are 27, 7, and 6, for C1, C2, and C3 covariates, respectively.

Selection Bias

Table 5.26 shows the performance of the evaluated designs in regards to selection bias. While the selection

	Selection Bias		
	Min	Mean	Max
CRD	0.00	0.00	0.00
PBD	16.65	24.66	35.05
ERADE.Neyman	39.04	45.97	58.60
ERADE.RSIHR	37.06	45.22	49.57
ERADE.Urn	42.77	52.60	69.26
ERADE.Baldi	53.58	63.86	77.91
DBCD.Neyman	5.32	12.66	48.68
DBCD.RSIHR	4.18	9.89	22.17
DBCD.Urn	9.34	23.80	59.18
DBCD.Baldi	28.22	49.27	81.61
SMLE.Neyman	1.75	6.77	42.03
SMLE.RSIHR	1.52	4.13	12.76
SMLE.Urn	4.80	16.87	58.12
SMLE.Baldi	38.75	46.84	57.74
EW1995.Neyman	2.93	10.03	42.99
EW1995.RSIHR	3.23	7.33	18.82
EW1995.Urn	6.65	22.08	171.28
EW1995.Baldi	35.03	47.57	67.13

Table 5.26: AZT trial reassessment with reduced sample size: selection bias under $H_1 : \beta_1 = 1.330$ for various designs in 10,000 simulated trials evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n =$ n needed for 90% power.

bias values have decreased due to the reduced sample sizes, the relative performances of the designs to each other remains consistent: CRD's selection bias is consistent at 0 by definition. EW1995.Urn again had the worst-case selection bias of 171.28. The lowest selection bias after CRD is provided by SMLE.RSIHR, with a minimum value of 1.52, the lowest average of 4.13, and the lowest maximum of 12.76, showing to be less predictable than PBD. RSIHR targeted by other designs also did well. ERADE again showed higher than average predictability, with average selection bias ranging from 45 to 64.

Expected Number of Failures

Table 5.27 displays the expected number of failures when reducing the sample size to be just sufficient for 90% power.

The range of the number of failures is 7 under ERADE.Urn, to 59 under ERADE.Neyman. ERADE.Urn yielded the least expected number of failures at 26.86, and EW1995.Neyman yielded the highest number of failures of 36.99. It also nearly had the highest maximum number of failures at 58, second only to ERADE.Neyman design. The poor performance of Neyman design, resulting in approximately 30% of enrolled subjects experiencing vertical transmission regardless of RAR design, is not surprising due to its tendency to place more subjects in the inferior arm (the control arm C, in this case) (See Table 5.23).

While indeed the absolute value of the expected number of failures has decreased since the original analysis due to decreased sample size, closer inspection yields that the range in the percentage of subjects experiencing failure has widened. For example, under CRD, the expected number of failures ranges from

	Expected Number of Failures		
	Min	Mean	Max
CRD	16.00	33.01	52.00
PBD	16.00	33.04	52.00
ERADE.Neyman	17.00	36.03	59.00
ERADE.RSIHR	14.00	31.25	53.00
ERADE.Urn	7.00	26.86	49.00
ERADE.Baldi	11.00	29.83	47.00
DBCD.Neyman	18.00	36.54	57.00
DBCD.RSIHR	14.00	31.72	50.00
DBCD.Urn	10.00	27.26	51.00
DBCD.Baldi	8.00	30.67	50.00
SMLE.Neyman	15.00	36.54	55.00
SMLE.RSIHR	16.00	32.07	53.00
SMLE.Urn	10.00	27.96	55.00
SMLE.Baldi	14.00	30.64	50.00
EW1995.Neyman	16.00	36.99	58.00
EW1995.RSIHR	14.00	32.29	54.00
EW1995.Urn	8.00	28.25	51.00
EW1995.Baldi	13.00	29.80	48.00

Table 5.27: AZT trial reassessment with reduced sample size: expected number of failures under $H_1 : \beta_1 = 1.330$ for various designs across 10,000 simulated trials evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n =$ n needed for 90% power.

16 to 52 (8% to 27% of 195 subjects), compared to 48 to 110 (10% to 13% of 477 subjects)of the original analysis. The largest range difference is seen in DBCD.Baldi, with percentage of subjects experiencing vertical transmission ranging from 8.18 to 18.66 (range of 10.48) in the original analysis, and widening to 3.54 to 22.12 (range of 18.58) in the reduced sample size analysis. The smallest increase in range of percentage of failures was 2.71 for EW1995.Baldi, with 7.55% to 20.75% experiencing failures in the original analyses' simulated trials, to 5.91% to 21.82% experiencing failures in the reduced sample size analyses' simulated trials. The increased range in percentage of failures when reducing sample size is not surprising, since the reduced sample size is associated with greater variability in the sequential estimation of parameters needed to calculate the probability of assignment to the AZT arm in the RAR designs.

Bias

Table 5.28 displays bias under the null hypothesis across 10,000 simulated trials.

Reduced sample size has resulted in the magnitude of bias under the null hypothesis increasing. For example, the range of bias in the original analysis with $n = 477$ was -0.740 to 0.729 (See Table 5.12). With sample size reduced to 195 for CRD, the range of bias was -1.344 to 1.430. For ERADE.Baldi design, the average bias was -0.014 when sample size was $n = 477$, with a minimum and maximum bias of -1.054 and 0.841. With a reduced sample size of 218, the average bias for ERADE.Baldi design was -0.017, with a minimum and maximum bias of -1.461 and 1.326, respectively. On average, ERADE.RSIHR and CRD

	Bias $E(\hat{\beta_1} - \beta_1)$		
	Min	Mean	Max
CRD	-1.344	0.000	1.430
PBD	-1.562	-0.001	1.442
ERADE.Neyman	-17.545	0.002	1.822
ERADE.RSIHR	-1.376	0.000	1.357
ERADE.Urn	-1.461	-0.002	1.326
ERADE.Baldi	-1.641	-0.017	1.540
DBCD.Neyman	-1.721	0.004	1.556
DBCD.RSIHR	-1.439	0.000	1.275
DBCD.Urn	-1.324	-0.002	1.229
DBCD.Baldi	-1.857	-0.020	1.305
SMLE.Neyman	-1.975	-0.002	2.067
SMLE.RSIHR	-1.251	-0.002	1.408
SMLE.Urn	-1.301	-0.004	1.261
SMLE.Baldi	-1.669	-0.017	1.279
EW1995.Neyman	-1.941	0.002	1.332
EW1995.RSIHR	-1.220	-0.001	1.390
EW1995.Urn	-1.448	0.002	1.104
EW1995.Baldi	-1.502	-0.019	1.263

Table 5.28: AZT trial reassessment with reduced sample size: bias of the treatment effect estimate $(E(\hat{\beta_1} - \beta_1))$ under $H_0 : \beta_1 = 0$ for various designs across 10,000 simulated trials evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n =$ n needed for 90% power.

perform best in the reduced sample size analysis, with an average bias of 0.0004 and 0.0005 (shown as 0.000 in Table 5.28), respectively. ERADE.Neyman has the worst magnitude of bias in the negative direction of -17.545. SMLE.Neyman has the worst magnitude of bias in the positive direction of 2.067. The largest bias on average in magnitude was seen in EW1995.Baldi, with an average bias of -0.019, followed closely by ERADE.Baldi with an average bias of -0.017. In the positive direction, Neyman allocation targeted by DBCD had the largest average bias of 0.004.

Relative Bias

Table 5.29 shows relative bias across 10,000 simulated trials for the designs evaluated.

Whilst the original analysis with $n = 477$ had minimum relative biases always greater than -100, meaning none of the designs concluded that the placebo arm had a higher probability of success, with a reduced sample size, that simulated guarantee is no longer in place. In fact, in their worst-case scenarios, each of the designs with reduced sample sizes led to a belief that the placebo arm had a higher probability of success. PBD had the lowest magnitude for minimum relative bias of -102.61.

DBCD.Baldi had the lowest average relative bias of 1.32, followed closely by EW1995.Baldi with an average of 1.77, and ERADE.Baldi, with an average of 1.87. CRD had an average relative bias of 3.91. Note that CRD, PBD, all RAR's targeting Neyman allocation, and most designs targeting RSIHR allocation, had very high maximum relative biases exceeding 1000. While PBD had the lowest magnitude of minimum

	Relative Bias $E(\frac{\hat{\beta}_1-\beta_1}{\beta_1}) \times 100$		
	Min	Mean	Max
CRD	-112.19	3.91	1338.07
PBD	-102.61	4.15	1352.95
ERADE.Neyman	-119.65	47.33	1250.22
ERADE.RSIHR	-124.23	4.12	1391.40
ERADE.Urn	-108.36	3.83	250.67
ERADE.Baldi	-186.26	1.87	172.48
DBCD.Neyman	-111.67	41.10	1254.49
DBCD.RSIHR	-128.02	3.13	206.62
DBCD.Urn	-112.90	3.49	175.56
DBCD.Baldi	-143.87	1.32	163.23
SMLE.Neyman	-105.39	25.62	1287.64
SMLE.RSIHR	-122.24	3.58	193.26
SMLE.Urn	-105.55	3.59	157.48
SMLE.Baldi	-131.97	2.19	150.51
EW1995.Neyman	-116.66	45.20	1248.72
EW1995.RSIHR	-110.15	3.28	195.59
EW1995.Urn	-103.01	3.12	153.63
EW1995.Baldi	-129.25	1.77	147.13

Table 5.29: AZT trial reassessment with reduced sample size: relative bias of the treatment effect estimate ($E(\frac{\hat{\beta}_1-\beta_1}{\beta_1}) \times 100$) under $H_1 : \beta_1 = 1.330$ for various designs across 10,000 simulated trials evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n =$ n needed for 90% power.

relative bias, it also had the largest magnitude of maximum relative bias, at 1352.95. This aligns with an estimate of $\hat{\beta}_1 = 3.130$, whilst the true $\beta_1 = 1.330146$, indicating an overestimation of the effect of the AZT treatment.

Type I error and Power

Table 5.30 display Type I error and power. Note how power always exceeds 90%, which is expected since the sample size necessary to achieve 90% power had been deduced through simulation and was passed through to this round of analysis. The interesting thing here that was hard to simulate was the ability of each design to control Type I Error.

We can compare with Table 5.16 and see that with the larger sample size of $n = 477$, CRD controlled Type I error at 0.0493, whilst reducing its sample size to 195 had led to loss of this control and a Type I error rate of 0.0510. Interestingly, Type I error for PBD moved in the opposite direction, decreasing from 0.0502 to 0.0477. Amongst the RAR designs, reducing the sample size had led to Type I error rates greater than 0.05 for ERADE.Baldi, DBCD.Neyman, DBCD.RSIHR, DBCD.Baldi, EW1995.Neyman, and EW1995.Urn. The other RAR designs were able to control Type I error below 0.05. Note that many of the RAR designs actually witnessed an improvement in Type I error rate relative to that yielded by the larger study. For example, ERADE.Baldi's Type I error decreased from 0.0513 to 0.0504, and SMLE.RSIHR's value decreased from 0.0525 to 0.0475. Surprisingly, the proportion of average bias in the reduced sample size analysis relative to

	Type I error	Power
CRD	0.0510	0.9073
PBD	0.0477	0.9065
ERADE.Neyman	0.0490	0.9001
ERADE.RSIHR	0.0474	0.9007
ERADE.Urn	0.0472	0.9028
ERADE.Baldi	0.0504	0.9002
DBCD.Neyman	0.0518	0.9036
DBCD.RSIHR	0.0526	0.9003
DBCD.Urn	0.0495	0.9038
DBCD.Baldi	0.0511	0.9051
SMLE.Neyman	0.0484	0.9039
SMLE.RSIHR	0.0475	0.9034
SMLE.Urn	0.0479	0.9032
SMLE.Baldi	0.0483	0.9090
EW1995.Neyman	0.0521	0.9051
EW1995.RSIHR	0.0448	0.9105
EW1995.Urn	0.0512	0.9044
EW1995.Baldi	0.0480	0.9010

Table 5.30: AZT trial reassessment with reduced sample size: Type I error and power for various designs across 10,000 simulated trials evaluating $\boldsymbol{p} = (0.917, 0.745)$ and $n =$ n needed for 90% power.

the average bias in the full sample size analysis was not predictive of the change in direction of Type I error. This speaks to the FDA's requirements of using simulation to show how Type I error is expected to behave in RAR trials.

Individual Desirability Function Definitions

The approach to defining individual desirability functions for the 11 design components discussed in Chapter 4 are similar in this reduced sample size analysis. For example, to determine the individual desirability function for bias under the null hypothesis, we anchor a value of 0 bias and map it to the individual desirability score of $d = 0$. We also see that from the simulated studies that the maximum witnessed bias was 2.0669, which we believe to be far too large. A maximum threshold of 0.5 is set, and anything above that should receive an individual desirability score of 0. The negative bias values are mapped to scores below 1, determined using the (1/11*1, 1/11*2, ...1/11*5)th percentiles of bias witnessed in the simulated trials. The positive bias values are mapped to scores above 1, determined using the (1/11*6, 1/11*7..., 1/11*9)th percentiles of bias witnessed in the simulated trials. (The 1/11*10 and 1/11*11th percentiles were greater than the threshold of 0.5). The thought process behind decision-making for the other nominal-the-better components are similar and will not be described in detail. Table 5.31 summarizes the selected values that map to specific individual desirability scores for NTB components treatment group size imbalance, bias, and relative bias.

Percentiles also aided in shaping individual desirability functions for larger-the-better (LTB) and smaller-the-better (STB) design components. For example, although Type I errors greater than 0.05 might be

	Nominal-the-Better (NTB) Components		
Individual Desirability Score d	Treatment Group Size Imbalance	Bias	Relative Bias
0.0	-100	-0.455	-186.261
0.2	-55	-0.308	-36.345
0.4	-40	-0.205	-22.483
0.6	-27	-0.119	-12.027
0.8	-7	-0.041	-2.529
1.0	0	0.000	0.000
0.8	11	0.037	6.699
0.6	61	0.115	16.677
0.4	84	0.200	28.918
0.2	101	0.301	47.391
0.0	120	0.500	1391.402

Table 5.31: AZT trial reassessment with reduced sample size: mapping definitions for individual desirability scores for nominal-the-better (NTB) design components.

immediately given a score of 0, there is some value in differentiating between Type I errors of 0.0504, which are very close to the 0.05 nominal level, versus Type I errors greater than 0.051. It is also helpful to give higher rewards Type I errors further away from and less than 0.05. Thus, errors greater than 0.051 are immediately given a score of 0, and the score of 0.0504 (the largest simulated Type I error smaller than 0.0505) is assigned an individual desirability score of 0.8. The remaining values are determined with the (1/4*1, 1/4*2, 1/4*3, 1/4*4)th percentiles, resulting in Type I error of 0.0486 being mapped to a score of 0.4, 0.0480 to a score of 0.6 0.0475 to 0.8, and 0.0448 to a score of 1. Similarly, the individual desirability functions for covariate imbalance for C1, C2, and C3 are defined mapping the the (1/5*0, 1/5*1, 1/5*2, 1/5*3, 1/5*4, and 1/5*5)th percentiles of the observed simulated probabilities of imbalance to individual desirability scores of (1, 0.8, 0.6, 0.4, 0.2, 0), respectively. The thought process behind decision-making for the other LTB and STB components are similar and will not be described in detail. Table 5.32 displays the selected values that map to specific individual desirability scores for LTB (power) and STB (accidental bias, covariate imbalance, selection bias, expected number of failures, Type I error, and n needed)components.

Overall Desirability Score

The weights for treatment group size imbalance, expected number of failures, covariate imbalance C1, covariate imbalance C2, covariate imbalance C3, selection bias, accidental bias factor estimate, bias under the null hypothesis, relative bias under the alternative hypothesis, Type I error, power, and sample size needed to obtain 90% power are denoted in the vector

$$\boldsymbol{w} = (w_{\text{imbal}}, w_{\text{fails}}, w_{\text{c1}}, w_{\text{c2}}, w_{\text{c3}}, w_{\text{sb}}, w_{\text{accbias}}, w_{\text{bias}}, w_{\text{relbias}}, w_{\text{alpha}}, w_{\text{power}}, w_{\text{n}}).$$

	Smaller-the-Better (STB) & Larger-the-Better (LTB) Components								
Individual Desirability Score d	Accidental Bias	Imbalance C1	Imbalance C2	Imbalance C3	Selection Bias	Expected No. of Failures	Type I Error	Power	n Needed
0	3.487	0.060	0.492	0.222	171.279	59.000	0.0505	0.8999	230
0.2	0.138	0.048	0.354	0.176	47.532	38.000	0.0504	0.900	220
0.4	0.078	0.037	0.240	0.153	33.419	34.000	0.0486	0.901	210
0.6	0.036	0.035	0.162	0.139	14.470	32.000	0.0480	0.904	200
0.8	0.035	0.031	0.110	0.132	7.332	29.000	0.0475	0.905	195
1	0.025	0.000	0.039	0.101	0.000	26.000	0.0448	0.910	189

Table 5.32: AZT trial reassessment with reduced sample size: mapping definitions for individual desirability scores for smaller-the-better (STB) and larger-the-better (LTB) design components.

We use the same set of weights from the original full sample size analysis that had resulted in selection of the Baldi allocation targeted by ERADE, and add a weight of 2 to sample size needed for 90% power. Specifically, we set $\boldsymbol{w} = (w_{\text{imbal}} = 0, w_{\text{fails}} = 3, w_{\text{c1}} = 0, w_{\text{c2}} = 1, w_{\text{c3}} = 1, w_{\text{sb}} = 0, w_{\text{accbias}} = 3, w_{\text{bias}} = 0.5, w_{\text{relbias}} = 2, w_{\text{alpha}} = 3, w_{\text{power}} = 1, w_{\text{n}} = 2)$. Tables 5.33 - 5.35 reveal the resulting mean overall desirability scores and probability of overall desirability score equaling zero for each of the designs.

	CRD	PBD	ERADE.Neyman	ERADE.RSIHR	ERADE.Urn	ERADE.Baldi	weight
Treatment Group Size Imbalance	0.814	0.967	0.369	0.834	0.244	0.405	0.000
Expected No. of Failures	0.533	0.530	0.386	0.627	0.823	0.701	0.182
Covariate Imbalance							
C1 (N(0,1))	1.000	1.000	0.347	0.758	0.066	0.513	0.000
C2 (linear time trend)	0.574	1.000	0.361	0.960	0.000	0.838	0.061
C3 (autocorrelated)	0.720	0.802	0.461	1.000	0.000	0.836	0.061
Selection Bias	1.000	0.492	0.224	0.233	0.192	0.174	0.000
Accidental Bias	0.807	0.625	0.378	0.541	0.216	0.656	0.182
Bias	0.380	0.379	0.374	0.378	0.392	0.366	0.030
Relative Bias	0.479	0.480	0.442	0.478	0.485	0.493	0.121
Type I Error	0.000	0.705	0.351	0.806	0.821	0.200	0.182
Power	0.881	0.852	0.200	0.289	0.520	0.215	0.061
N Needed for 90% Power	0.800	0.800	0.967	0.967	0.400	0.240	0.121
Overall Desirability D (mean)	0.000	0.601	0.376	0.607	0.000	0.407	
Prob(Overall Desirability D = 0)	1.000	0.156	0.165	0.158	1.000	0.179	

Table 5.33: AZT trial reassessment with reduced sample size: mean individual desirability scores for 12 considered design characteristics, mean overall desirability Score, and probability that overall desirability score is 0, for CRD, PBD, ERADE.Neyman, ERADE.RSIHR, ERADE.Urn, and ERADE.Baldi designs.

	DBCD.Neyman	DBCD.RSIHR	DBCD.Urn	DBCD.Baldi	SMLE.Neyman	SMLE.RSIHR	weight
Treatment Group Size Imbalance	0.374	0.840	0.207	0.356	0.424	0.812	0.000
Expected No. of Failures	0.366	0.603	0.804	0.658	0.366	0.583	0.182
Covariate Imbalance							
C1 (N(0,1))	0.373	0.804	0.000	0.438	0.656	0.809	0.000
C2 (linear time trend)	0.378	0.825	0.052	0.720	0.408	0.597	0.061
C3 (autocorrelated)	0.358	0.784	0.117	0.543	0.399	0.687	0.061
Selection Bias	0.665	0.729	0.501	0.216	0.817	0.887	0.000
Accidental Bias	0.383	0.639	0.200	0.790	0.346	0.728	0.182
Bias	0.376	0.375	0.395	0.366	0.375	0.380	0.030
Relative Bias	0.440	0.482	0.485	0.496	0.456	0.477	0.121
Type I Error	0.000	0.000	0.297	0.000	0.450	0.789	0.182
Power	0.591	0.230	0.614	0.800	0.629	0.573	0.061
N Needed for 90% Power	0.867	0.867	0.300	0.080	0.800	0.800	0.121
Overall Desirability D (mean)	0.000	0.000	0.301	0.000	0.403	0.606	
Prob(Overall Desirability D = 0)	1.000	1.000	0.139	1.000	0.160	0.154	

Table 5.34: AZT trial reassessment with reduced sample size: mean individual desirability scores for 12 considered design characteristics, mean overall desirability Score, and probability that overall desirability score is 0, for DBCD.Neyman DBCD.RSIHR, DBCD.Urn, DBCD.Baldi, SMLE.Neyman, and SMLE.RSIHR designs.

	SMLE.Urn	SMLE.Baldi	EW1995.Neyman	EW1995.RSIHR	EW1995.Urn	EW1995.Baldi	weight
Treatment Group Size Imbalance	0.262	0.362	0.365	0.836	0.221	0.367	0.000
Expected No. of Failures	0.771	0.658	0.350	0.570	0.761	0.701	0.182
Covariate Imbalance							
C1 (N(0,1))	0.117	0.566	0.340	0.753	0.087	0.513	0.000
C2 (linear time trend)	0.113	0.556	0.357	0.749	0.093	0.617	0.061
C3 (autocorrelated)	0.149	0.590	0.400	0.804	0.150	0.371	0.061
Selection Bias	0.603	0.218	0.730	0.800	0.524	0.217	0.000
Accidental Bias	0.191	0.759	0.352	0.800	0.195	0.718	0.182
Bias	0.394	0.368	0.378	0.381	0.402	0.368	0.030
Relative Bias	0.489	0.499	0.446	0.484	0.488	0.494	0.121
Type I Error	0.621	0.483	0.000	1.000	0.000	0.583	0.182
Power	0.556	0.944	0.800	1.000	0.700	0.333	0.061
N Needed for 90% Power	0.300	0.100	0.800	0.760	0.200	0.200	0.121
Overall Desirability D (mean)	0.358	0.453	0.000	0.675	0.000	0.470	
Prob(Overall Desirability D = 0)	0.139	0.169	1.000	0.159	1.000	0.176	

Table 5.35: AZT trial reassessment with reduced sample size: mean individual desirability scores for 12 considered design characteristics, mean overall desirability Score, and probability that overall desirability score is 0, for SMLE.Urn, SMLE.Baldi, EW1995.Neyman, EW1995.RSIHR, EW1995.Urn, and EW1995.Baldi designs

After reducing the sample size to be just sufficient for 90% power and using the weights as defined, notice how several designs yield a score of 0. This is due to a more stringent treatment of scoring for the Type I error component. Baldi as targeted by SMLE is no longer the highest overall scorer. With the reduced sample sizes, we see that RSIHR allocation outperforms the others, especially when targeted by EW1995, with an overall desirability score of $D = 0.675$, and probability of overall D being 0 at 0.159. SMLE.RSIHR and ERADE.RSIHR also do well when compared to the other designs assessed, with overall desirability scores of approximately $D = 0.60$. Note that DBCD.RSIHR's ability to control Type I error has driven its overall desirability score to $D = 0$. The higher score of EW1995.RSIHR as compared to SMLE.RSIHR and ERADE.RSIHR is mainly due to the higher scores of components Type I error and Power, since EW1995.RSIHR scored perfectly in regards to those two components. PBD also does well, with similar performance to RSIHR allocation not targeted by EW1995, with an overall desirability score of $D = 0.601$ and probability of overall desirability being 0 at 0.156. SMLE.Baldi no longer scores well with reduced sample size, precisely due to its requirement for a sample size of 225 subjects, second only to DBCD.Baldi requiring 226. With the number of subjects needed accounting for more than 10% of the overall score (normalized weight = 0.121), the overall score for SMLE.Baldi was penalized sufficiently to lose to RSIHR allocation. For example, EW1995.RSIHR required a sample size of 196, and ERADE.RSIHR only required a sample size of 190 (less than that required by CRD).

We reassess the designs with the other weights evaluated during the original full sample size analysis. We add weights to treatment group size imbalance and covariate imbalance for covariate C1. Specifically, let the weight vector be $\boldsymbol{w} = (w_{\text{imbal}} = 0.5, w_{\text{fails}} = 3, w_{\text{c1}} = 2, w_{\text{c2}} = 1, w_{\text{c3}} = 1, w_{\text{sb}} = 0, w_{\text{accbias}} = 3, w_{\text{bias}} = 0.5, w_{\text{relbias}} = 2, w_{\text{alpha}} = 3, w_{\text{power}} = 1, w_{\text{n}} = 2)$. Recall that this weight vector, without the n needed component, resulted in selection of PBD during the full sample size analysis. The analysis is repeated with the reduced sample size, and places high importance on sample size needed to attain 90% power. The detailed results are not displayed for brevity. To summarize, PBD still performs well, with an overall desirability score of 0.642 and a probability of overall desirability being 0 of 0.156. Note CRD, which had done well with this set of weights when $n = 477$, had an overall desirability score of 0, penalized due to its inability to control Type I error at the 5% level. EW1995.RSIHR outperforms PBD and the other designs, with an overall desirability score of 0.686, and a probability of overall desirability being 0 of 0.159. SMLE.RSIHR and ERADE.RSIHR also score on average greater than 0.6; DBCD.RSIHR again has an average score of 0 due to its higher Type I error. Baldi allocations score in the range of 0.416 to 0.469, except for when utilizing DBCD, in which Type I error pulls the overall desirability score to 0.

An analysis with the third of weights, letting $\boldsymbol{w} = (w_{\text{imbal}} = 1, w_{\text{fails}} = 3, w_{\text{c1}} = 2, w_{\text{c2}} = 0, w_{\text{c3}} = 0, w_{\text{sb}} = 1, w_{\text{accbias}} = 2, w_{\text{bias}} = 0.5, w_{\text{relbias}} = 3, w_{\text{alpha}} = 3, w_{\text{power}} = 2, w_{\text{n}} = 2)$ is repeated with the reduced sample size scenario. Recall when $n = 477$, the CRD design yielded the highest overall desirability score with this set of weights. The performance of EW1995.RSIHR is again highest relative to the other designs, with an overall desirability score of 0.671 and probability of overall desirability being 0 of 0.159. RSIHR allocation targeted by ERADE and SMLE, and PBD again perform well.

In conclusion, RSIHR allocation targeted by EW1995 performed best with individual desirability functions defined in Tables 5.31 and 5.32, and weights defined in Tables 5.33 to 5.35. The other weight vectors utilized in the original full sample size analysis consistently pointed to EW1995.RSIHR as the highest scoring design. One of the main differentiations for this design was its ability to attain 90% power with relatively fewer subjects, $n = 190$, specifically. This is less than 40% of the sample size of 477 during the interim analysis of the original trial after which the trial was stopped early for overwhelming efficacy.

5.4 Discussion

In the redesigning of an HIV clinical trial evaluating the efficacy of zidovudine (AZT) in the prevention of vertical HIV transmission, a preliminary analysis utilizing the full sample size of $n = 477$ subjects concluded that among 18 evaluated designs, permuted block design with a block size of 8 had the overall highest quality with respect to the design components considered. In a sensitivity analysis, we found that placing less weight on treatment group size imbalance led to the favoring of the DBCD design targeting Baldi allocation, yet placing more weight on controlling selection bias and probability of covariate imbalance for covariate type C1 resulted in the favoring of the design actually used in the trial: complete randomized design. One quality of the original trial was that it was highly overpowered, aiding in early termination at the end of a first interim look at the data in favor of zidovudine (AZT) treatment. Two potential factors could be contributors to the overpowering, which equated to enrolling more subjects than necessary to decisively reject the null hypothesis of no treatment effect. The first is an underestimation of the effect size, and the second is an underestimation of the proportions of pregnancies which would be considered evaluable for the trial.

When reducing the sample size to one just sufficient for 90% power, the design quality assessment framework identified the lack of control of Type I error as a dealbreaker for several of the assessed designs, and concluded that RSIHR allocation targeted by the EW1995 design resulted in the design with the overall best quality. This selection proved robust to the weights inspected in the preliminary analysis with the full sample size.

The evaluation of a clinical trial as assessed by desirability functions is sensitive to the definition of individual desirability functions that score the design components of interest (as demonstrated in Section 4.3.3, where two different definitions of the individual desirability function for total response were considered), and the weights of these design components in the final scoring of the design (as demonstrated in this chapter). The sensitivity of design selection to weights of design components underscores the importance of identifying the design components important to a design's quality, and understanding the true preferences of stakeholders and accurately reflecting them in the scoring process of each design.

Epilogue

The evaluation of a clinical trial using desirability functions is sensitive to both the definition of individual desirability functions that score the design components of interest, and the weights of these design components in the final scoring of the design. Sensitivity analysis for weights was discussed in Chapter 5, underscoring the importance of understanding the true preferences of stakeholders and accurately reflecting them in the scoring process of each design.

Although this work focused primarily on a subset of clinical trial design characteristics that contribute to a trial's overall quality, the framework presented in Chapter 4 is flexible to the needs of the clinical trial stakeholders. There are many design components that can be included in the overall desirability function of a design. For example, in this work, we did not discuss components with binary values. The design evaluation framework can potentially include logistical aspects of a design such as time to completion, financial cost, and ease of implementation. Future expansion on the framework can also include design formats other than randomization sequences – one could simulate results of trials with more than two treatment arms, interim analyses with various stopping rules, platform designs where treatment arms can be dropped or added throughout the trial, wedge designs which allow random and sequential crossover of clusters from one treatment arm to another until all subjects have been exposed to the treatment, clustered randomized trials, and many others. The framework can also be extended towards not just selection of a design, but selection of a primary endpoint. For example, in deciding whether a primary endpoint should be binary (e.g. "proportion of subjects in each treatment arm achieving at least a 30% reduction in intact parathyroid hormone (iPTH)") or continuous (e.g. "mean percent change in intact parathyroid hormone (iPTH) from baseline to week 26"), desirability scores could be calculated for the designs evaluating these endpoints, with significant weight given to the power component of the design.

The applications of the desirability function framework in this dissertation confirmed some strengths and weaknesses of the evaluated designs. Complete Randomized Design (CRD)'s strengths are: a) randomness of treatment assignment resulting in zero selection bias; b) low probability of normally distributed covariate imbalance; c) fewer subjects needed to attain desired powerer; and d) when sample size is sufficiently large, treatment group size imbalance is usually small; small treatment group size imbalances are an important factor in controlling against accidental bias. CRD's weaknesses include: a) inability to prevent time trends in responses or in covariates (covariate C2), resulting in potential misinterpretation of the treatment effect; and b) exclusion of ethical considerations such as minimizing the expected number or failures in the trial. Forced balance designs and biased coin designs perform well in: a) balancing the size of the two groups in a trial; b) balancing covariates of different types (e.g. normally distributed or having a linear time trend); c) letting time trends influencing the response do so equally; d) and attaining desired power. Their weaknesses may include predictability of treatment assignments when the block size or current treatment group sizes can be guessed. There is a trade-off in forced balance designs between predictability due to small block sizes and stronger protection against chronological bias.

Response-adaptive randomization (RAR) designs aim to assign subjects to a trial in line with a specific target allocation. Variability in the estimate of the allocation proportion is an undesirable quality in RAR designs because the variability negatively decreases the power of the design. The variability of the probability of the next subject being assigned to a specific arm (φ) throughout a study is a component that can be included in the overall desirability function. The ERADE design has minimal variability in treatment assignment probability φ. Other RAR designs SMLE, DBCD, and EW1995 have varying levels depending on tuning parameters selected by the user. SMLE has shown consistently throughout this work to be the least predictable of the RAR designs considered. The target allocation targeted by RAR designs plays a big role on the design's performance with respect to the components discussed in this work. Designs targeting RSIHR and Urn allocations have fewer expected number of failures – their tendency to assign consecutive subjects to the better-performing arm leads to significantly less hedging to time trends in either the response variable or confounding covariates. Designs that target Neyman allocation have higher power than RAR designs that target other allocations, but also are more likely to have more extreme bias in the treatment effect estimate. RAR designs have shown varying performance relative to balancing covariates.

The comparison of designs using desirability functions in this work highlights some challenges of RAR designs: although RAR designs target an allocation scheme that often fulfills an inferential or ethical objective, the variability of their performances require us to study and understand their properties before implementation. We can investigate their theoretical properties or conduct a detailed simulation. Certain scenarios are well-suited for RAR designs; others are best studied with traditional designs like CRD and PBD. The contraceptive study of Application 1 in Chapter 4 included no time trends in the response variable, but still led to selection of the Random Block Design with blocks filled with Truncated Binomial Design. The scleroderma study of Application 2 in Chapter 4 incorporated a time trend, which led to the selection of Permuted Block Design. The scleroderma study of Application 3 in Chapter 4 removed the time trend and selected RSIHR2 allocation. In the AIDS application of Chapter 5, Permuted Block Design and Complete Randomized Design ranked highest amongst designs evaluated, until less weight was placed on protecting against covariate imbalance, which resulted in the selection of the Baldi allocation targeted by DBCD. The reduction of sample size in the AIDS application resulted in a robust performance and selection of RSIHR allocation targeted by EW1995 design in the presence of various weight settings. The varying performances of the designs in these different applications underline the importance of simulation when considering RAR designs. We recommend RAR designs when ethical objectives are an important part of a trial, and caution users to carefully consider RAR designs should time trends or limited financial or logistic resources be a concern. Baldi allocation is always worth considering as it is able to simultaneously incorporate inferential and ethical objectives.

The flexibility of the desirability function framework in assessing clinical trial design is highly attractive and can incorporate the preferences of stakeholders from different key functions. Weaknesses of the proposed desirability function framework in evaluation of clinical trial design include the subjective nature of defining individual desirability function shapes and design characteristic weights. Users of the framework should clearly understand their preferences prior to exploring simulation results, to avoid cherry-picking function shapes and weights that point to their pre-established preferences. Although the Delphi method discussed in Chapter 4 is recommended, it may be difficult in practice for stakeholders to come to a consensus of individual desirability function definitions and weights.

We encourage readers to explore our Shiny application at `https://priscillakyen.shinyapps.io/DesignEvaluation_beta/`. This online tool allows users to answer some questions about their clinical trial (e.g. is your outcome binary or continuous?, what level of Type I error can you accept?, what is your anticipated treatment effect size?, is there any reason to suspect time trends?) and to select design candidates. The website then simulates trials with these designs and asks users about their preferences regarding the design characteristics discussed in Chapter 4 (individual desirability function shape parameters) and their relative importance (weight parameters) before outputting overall desirability scores and the estimated probability that the overall desirability score of a design is zero. Visual plots help users understand how they have defined their individual desirability functions, and the interface is immediately reactive to weight specifications, so that simple sensitivity analyses on weights can be performed. The final scores can help users select the best design to fulfill their objectives. The website has limitations: it currently only includes two-arm trials and a limited number of randomization designs. Furthermore, the Shiny app tends to run slower through the R server than the same app runs locally on a laptop with 8GB installed memory (RAM), with a 64-bit operating system and an i7 processor @2.10 GHz. Also, the Shiny app only allows one user to access a single function within the program at a time. Given that a simulation comparing five trials could take over an hour (depending on the sample size), this is something that would be hard to scale up. In spite of its limitations, the website still is a valuable tool to help users understand the impact of design choice on design characteristics and potentially the overall interpretation of the results. This Shiny app could be more useful to the community if it could implement more outcome types (e.g. time-to-event) and clinical trial designs (e.g. platform designs, crossover designs, stratification factors, early stopping rules, etc.). Although there is room for improvement, the desirability framework presented in this work is now food for thought and is flexible enough to be modified to suit one's needs during clinical trial design selection.

Appendices

Appendix A

R Code

```
#FUNCTION THAT SOLVES FOR TOTAL SAMPLE SIZE NEEDED TO DETECT A SPECIFIED EFFECT SIZE (u1-u2) FOR A GIVEN TYPE I ERROR AND POWER
Solve.N <- function(u1, u2, var1, var2, alpha, power){
  n.e <- n.c <- (((qnorm(1-alpha/2) + qnorm(power))^2*(var1 + var2))/(u1-u2)^2)
  N = n.e + n.c
  resultvec <- c(N, u1, u2, var1, var2)
  return(resultvec)
}

#FUNCTION THAT SOLVES FOR R = nE/nC
Solve.R <- function(u1, u2, var1, var2, corr){
  cov12 <- corr*sqrt(var1)*sqrt(var2)
  f <- function(R) ( u1*var2*R^2 - u1*cov12*R^(3/2) + u2*cov12*sqrt(R) - u2*var1)
  R <- uniroot(f, lower=0.000001, upper=100000)$root
  result <- c(u1, u2, var1, var2, corr, R)
  return(result)
}
```

Appendix B

Website Manual

Introduction

A Shiny app utilizing the framework of Chapter 4 is available at `https://priscillakyen.shinyapps.io/DesignEvaluation_beta/`. The goal of this website is to allow users to simulate different clinical trial designs to assess their strengths and weaknesses, and to provide users a framework in which they can quantitatively define their preferences and select a design that fulfills their research objectives. We begin with some notation that is used throughout the manual, followed by Website Instructions, Summary of Designs, and Evaluated Characteristics of Designs. *Note that in this manual, blue text and Equation numbers can be clicked on to take you to relevant sections of the manual.*

Notation:

n_E	number of subjects in experimental arm (group E)
n_C	number of subjects in control arm (group C)
$n_E(j)$	number of subjects in experimental arm (group E) at the time of the j^{th} patient
$n_C(j)$	number of subjects in control arm (group C) at the time of the j^{th} patient
n	total sample size = $n_E + n_C$
Y_{E_j}	response of subject j in experimental arm (group E)
Y_{C_j}	response of subject j in control arm (group C)
f_E	number of failures in experimental arm (group E)
f_C	number of failures in control arm (group C)
α	alpha-level: probability of rejecting a null hypothesis when in fact the null hypothesis is true
T_j	experimental arm indicator variable for subject j ($j = 1, \ldots, n$): 1 if in experimental arm, 0 if in control arm
iter	the number of iterations performed in a simulation, with each iteration completing one trial
d_k	individual desirability score for the k^{th} value of a characteristic ($k = 1, \ldots, \text{iter}$)
L	lower limit of a desirability function
U	upper limit of a desirability function
T	target value of a desirablity function (NTB variables only)
r	scale parameter for individual desirability functions (STB and LTB only)
r_1, r_2	scale parameters for individual desirability functions (NTB only)
w_i	weight for characteristic i ($i = 1, \ldots, m$) in calculation of D
D	overall desirability score for a design

Website Instructions

In order to find the design best suited to your needs, the website guides you through 3 tabs:

Step 1. See design characteristics.

Step 2. Individual Desirability Functions.

Step 3. Overall Desirability Functions.

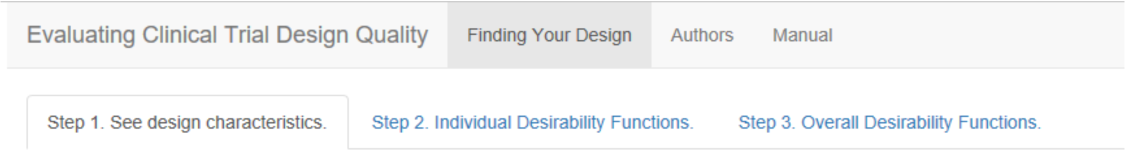

Evaluating Clinical Trial Design Quality
Finding Your Design
Authors
Manual
Step 1. See design characteristics.
Step 2. Individual Desirability Functions.
Step 3. Overall Desirability Functions.

Figure B.1

You can navigate between these tabs without losing your work. The goal of the first tab is for the program to understand what you *anticipate* or *expect* from the data, followed by an automated simulation that provides performance results of each design under evaluation with regards to certain design characteristics.

The goal of the second tab is to provide you with performance summaries of the various designs so that you can reflect on how you value different realizations of different characteristics. The program will walk you through how to shape individual desirability functions for each characteristic. If a certain characteristic is of zero interest to you, you can leave it at its default values, and tell the program in the third step that you do not care to include the performance of each design in regard to this characteristic in its final evaluation.

In the third and last tab, you are asked to rate how important each characteristic is to you, on a scale between 0 and 3. The overall desirability scores will then be calculated taking into account these ratings.

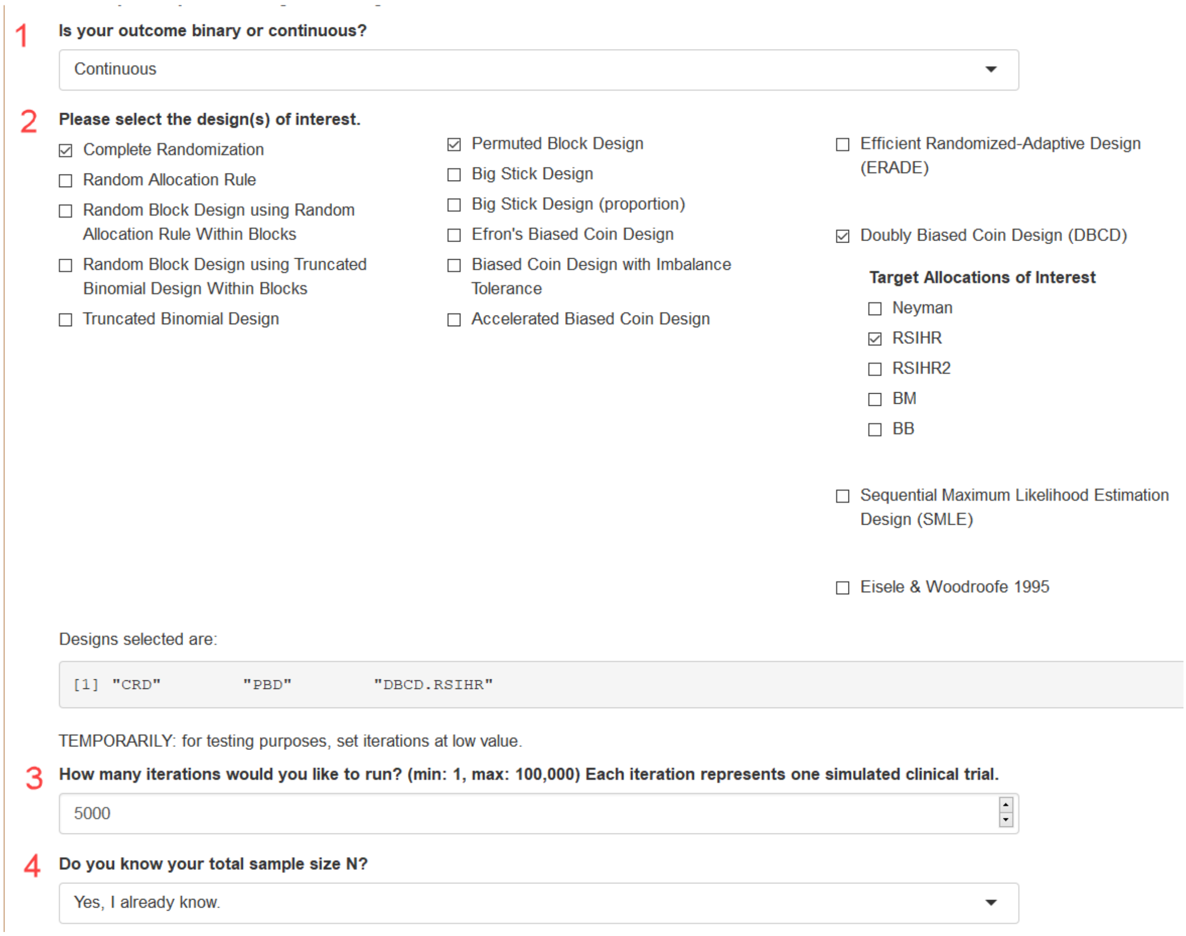
Step 1.

1. Is your outcome binary or continuous?
If your outcome can take on two values (Yes vs. No), select Binary.
If your outcome can take on a range of values, select Continuous. Note that the RAR designs on this website cannot handle negative values.

2. Please select the design(s) of interest.
See Summary of Designs for background on the designs listed.

The designs in the left two columns are not adaptive. The designs in the right column are response-adaptive randomization (RAR) designs. Upon clicking one of these RARs, you will be asked to select a target allocation (Neyman, RSIHR, Urn, and Baldi for binary responses, or Neyman, RSIHR, RSIHR2, BB, BM for continuous responses).

The short names of the designs you selected are displayed before the start of the next question. In this example, we have selected Complete Randomization, Permuted Block Design, and Doubly Biased Coin



1 **Is your outcome binary or continuous?**

Continuous ▾

2 **Please select the design(s) of interest.**

☑ Complete Randomization

☐ Random Allocation Rule

☐ Random Block Design using Random Allocation Rule Within Blocks

☐ Random Block Design using Truncated Binomial Design Within Blocks

☐ Truncated Binomial Design

☑ Permuted Block Design

☐ Big Stick Design

☐ Big Stick Design (proportion)

☐ Efron's Biased Coin Design

☐ Biased Coin Design with Imbalance Tolerance

☐ Accelerated Biased Coin Design

☐ Efficient Randomized-Adaptive Design (ERADE)

☑ Doubly Biased Coin Design (DBCD)

Target Allocations of Interest

☐ Neyman

☑ RSIHR

☐ RSIHR2

☐ BM

☐ BB

☐ Sequential Maximum Likelihood Estimation Design (SMLE)

☐ Eisele & Woodroofe 1995

Designs selected are:

```
[1] "CRD"        "PBD"        "DBCD.RSIHR"
```

TEMPORARILY: for testing purposes, set iterations at low value.

3 **How many iterations would you like to run? (min: 1, max: 100,000) Each iteration represents one simulated clinical trial.**

5000

4 **Do you know your total sample size N?**

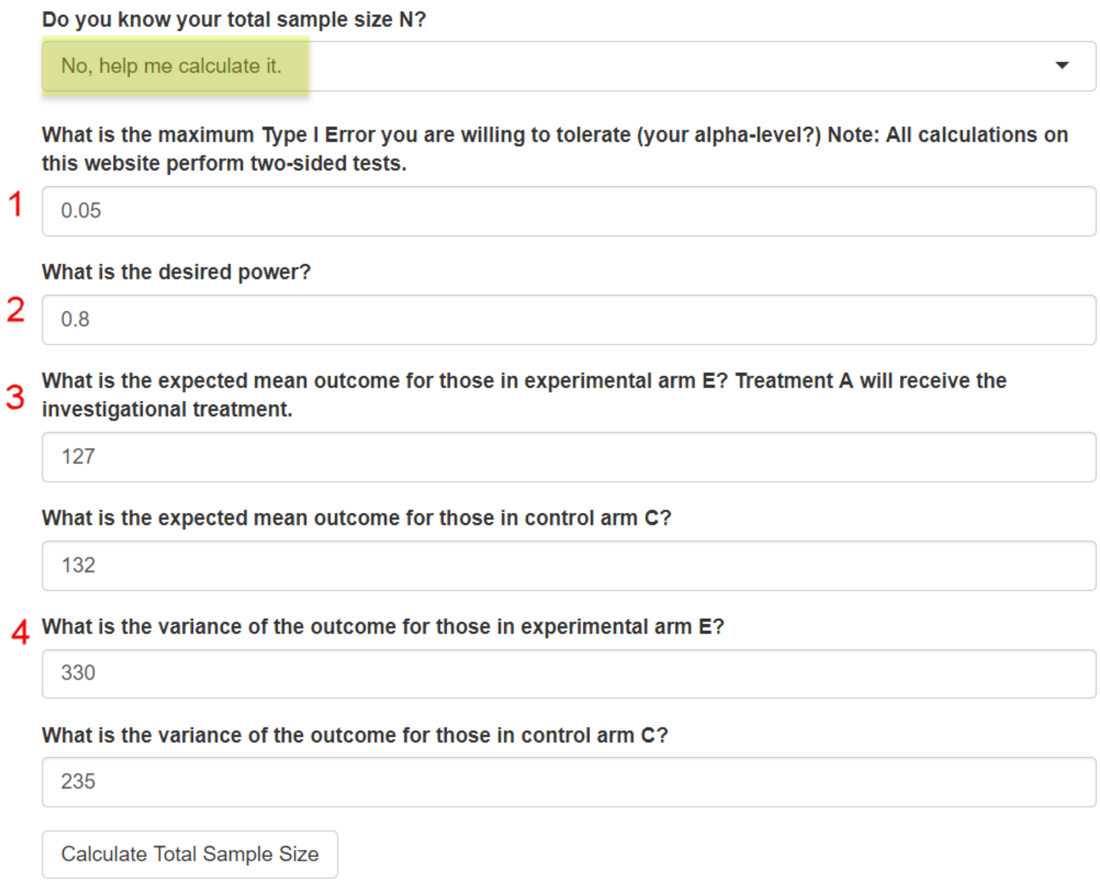
Yes, I already know. ▾

Figure B.2

Design targeting RSIHR allocation. The abbreviations for the designs are listed as “CRD”, “PBD”, and “DBCD.RSIHR”.

3. How many iterations would you like to run?
Each iteration of a simulation represents one simulated study. All the designs you selected will be evaluated in each iteration. If you are just setting up your study for the first time, use a small number of iterations (i.e. 10) for a preliminary look at results. When you are more certain about your simulation setup, you can set the number of iterations to a larger number (i.e. 1000 or 10000) to get more accurate estimates of characteristics that are measured as an average across all simulated studies. For example, the probability of covariate imbalance exceeding 0.3 is calculated as the number of iterations in which the design resulted in a covariate imbalance greater than 0.3 divided by the total number of iterations.

4. Do you know your total sample size N?
If you are unsure, say “No” and the program will walk you through how to calculate the required sample size for a specified Type I error and Power. Note that this calculation is for non-adaptive designs.



Do you know your total sample size N?
No, help me calculate it.

What is the maximum Type I Error you are willing to tolerate (your alpha-level?) Note: All calculations on this website perform two-sided tests.
0.05

What is the desired power?
0.8

What is the expected mean outcome for those in experimental arm E? Treatment A will receive the investigational treatment.
127

What is the expected mean outcome for those in control arm C?
132

What is the variance of the outcome for those in experimental arm E?
330

What is the variance of the outcome for those in control arm C?
235

Figure B.3

1. What is the maximum Type I error you are willing to tolerate (your alpha-level?) Note: All calculations on this website perform two-sided tests.

Type I error is the probability of rejecting the null hypothesis (e.g.: H_0: no treatment difference) in favor of the alternative hypothesis (e.g.: H_1: there is a treatment difference), when in fact the null hypothesis is true. Typically, controlling the Type I error at an alpha-level of 0.05 is standard. The lower the Type I error you are willing to accept, the more subjects you will need in your trial.

2. What is the desired power?
Power is the probability of rejecting the null hypothesis when the alternative hypothesis is actually true. Typically, power in the design stage of selecting a clinical trial design is set to at least 0.80. Do not request a power greater than 0.99. The higher the power you require, the more subjects you will need in your trial.

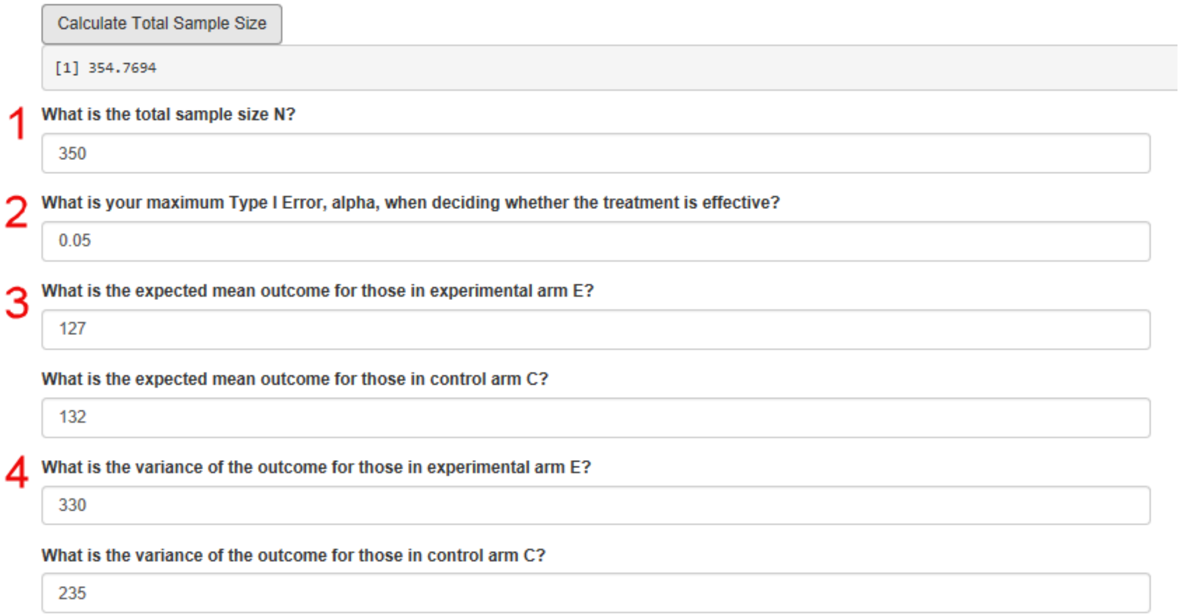
3. What is the mean outcome for those in experimental arm E? Treatment E will receive the investigational treatment.
This and the following question, which asks for the mean outcome for those in treatment arm C, ask what you expect the average responses in the experimental arm (E) and control arm (C) to be. These are μ_E and μ_C, respectively. The smaller the difference between these two means, the smaller the *effect size*, and thus the more subjects you will need in the trial.

4. What is the variance of the outcome for those in treatment E?
This and the following question ask how much variance you expect to see in the responses of the experimental arm (E) and the control arm (C). These are σ_E^2 and σ_C^2, respectively. The larger the variance, the more uncertainty there is when estimating the true mean of the outcome, and thus the more subjects you will need in the trial.

Let $Y_E \sim N(\mu_E, \sigma_E^2)$, $Y_C \sim N(\mu_C, \sigma_C^2)$; $Y_E \perp\!\!\!\perp Y_C$. When you click on “Calculate Total Sample Size”, the equation below is used to calculate the sample size needed.

$$n = n_E + n_C = 2\frac{(z_{1-\alpha/2} + z_{1-\tilde{\beta}})^2(\sigma_E^2 + \sigma_C^2)}{(\mu_E - \mu_C)^2}. \tag{B.1}$$

In this example, a total sample size of 355 subjects is needed.



```
[1] 354.7694
```

What is the total sample size N?

350

What is your maximum Type I Error, alpha, when deciding whether the treatment is effective?

0.05

What is the expected mean outcome for those in experimental arm E?

127

What is the expected mean outcome for those in control arm C?

132

What is the variance of the outcome for those in experimental arm E?

330

What is the variance of the outcome for those in control arm C?

235

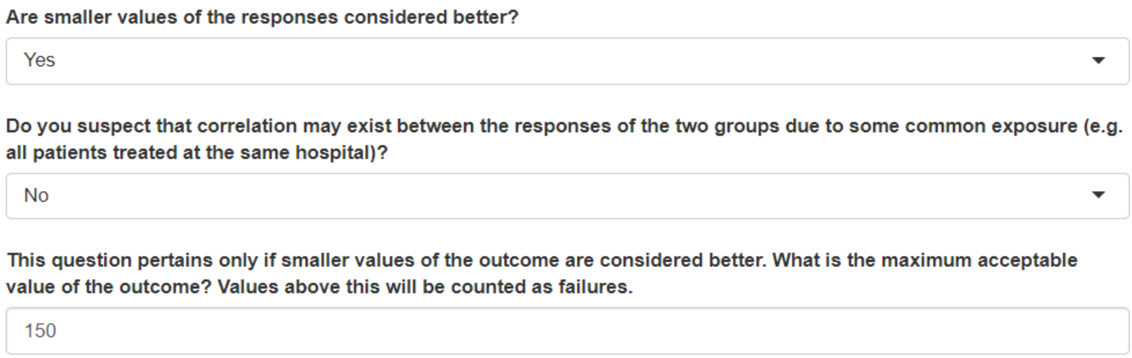
Figure B.4

1. What is the total sample size N?
Input the total sample size $n_E + n_C = n$ here.

2. What is the maximum Type I error, alpha, when deciding when the treatment is effective?
Type I error is the probability of rejecting the null hypothesis (e.g.: H_0: no treatment difference) in favor of the alternative hypothesis (e.g.: H_1: there is a treatment difference), when in fact the null hypothesis is true. Typically, controlling the Type I error at an alpha-level of 0.05 is standard. In this program, when the p-value of the coefficient of the treatment effect ($\hat{\beta}_1$) is less than the specified alpha, the null hypothesis is rejected.

3. What is the mean outcome for those in treatment E?
This and the following question, which asks for the mean outcome for those in treatment arm C, ask what you expect the average responses in the experimental arm (E) and control arm (B) to be. These are μ_E and μ_C, respectively.

4. What is the variance of the outcome for those in treatment E?
This and the following question ask how much variance you expect to see in the responses of the experimental arm (E) and the control arm (C). These are σ_E^2 and σ_C^2, respectively.



Are smaller values of the responses considered better?

Yes

Do you suspect that correlation may exist between the responses of the two groups due to some common exposure (e.g. all patients treated at the same hospital)?

No

This question pertains only if smaller values of the outcome are considered better. What is the maximum acceptable value of the outcome? Values above this will be counted as failures.

150

Figure B.5

1. Are smaller values of the response considered better?
The purpose of this question is so that the program may later know how to evaluate the total response $\bar{Y}_E n_E + \bar{Y}_C n_C$. Also, if you select "Yes" to this question, you will be further asked **2** below.

2. Do you suspect that correlation may exist between the responses of the two groups due to some common exposure (e.g. all patients treated at the same hospital?)
If you select "No", the program will generate independent responses.
If you select "Yes", the program will subsequently ask: "What is the level of correlation (range: (0,1]?)" Please input a value ranging between 0 and 1. This value can be estimated from pilot data or drawn from previous literature. The program will then generate correlated responses as follows:

> A set of outcomes of length n is simulated for *each* treatment group, where the outcome is simulated as such:
>
> - Independently generate a vector $Z_1 \sim N(0,1)$ of length n and a second vector $Z_2 \sim N(0,1)$ also of length n.
> - Let $Z_3 = \rho Z_1 + \sqrt{1-\rho^2} Z_2$.
> - Then n outcomes for treatment group E are calculated as $Y_E = \mu_E + \sqrt{\sigma_E^2} Z_1$, while the n outcomes for treatment group C are calculated as $Y_C = \mu_C + \sqrt{\sigma_C^2} Z_3$.

It will furthermore add to the list of designs you have selected for evaluation. If you have not selected any RAR designs, the program will automatically consider DBCD targeting R.corr. If you have selected any RAR designs, the program will target R.corr with the RAR designs you selected.

3. This question pertains only if smaller values of the outcome are considered better. What is the maximum acceptable value of the outcome? Values above this will be counted as failures.
When responses are considered "smaller-the-better", we ask for a maximum acceptable threshold of the value. Any responses above this threshold will be counted as a failure. The total number of failures in each treatment arm are summed and compared at the end of the analysis.

For the next several questions, consider the following. Let X be

$$\boldsymbol{X} = \begin{bmatrix} 1 & T_1 \\ 1 & T_2 \\ \vdots & \vdots \\ 1 & T_j \\ \vdots & \vdots \\ 1 & T_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}. \tag{B.2}$$

$$\boldsymbol{Z} = \begin{bmatrix} Z_{1\text{linEC}} & Z_{1\text{linE}} & Z_{1\text{linC}} & Z_{1\text{stepEC}} & Z_{1\text{stepE}} & Z_{1\text{stepC}} & Z_{1\text{logEC}} & Z_{1\text{logE}} & Z_{1\text{logC}} \\ Z_{2\text{linEC}} & Z_{2\text{linE}} & Z_{2\text{linC}} & Z_{2\text{stepEC}} & Z_{2\text{stepE}} & Z_{2\text{stepC}} & Z_{2\text{logEC}} & Z_{2\text{logE}} & Z_{2\text{logC}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{j\text{linEC}} & Z_{j\text{linE}} & Z_{j\text{linC}} & Z_{j\text{stepEC}} & Z_{j\text{stepE}} & Z_{j\text{stepC}} & Z_{j\text{logEC}} & Z_{j\text{logE}} & Z_{j\text{logC}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{n\text{linEC}} & Z_{n\text{linE}} & Z_{n\text{linC}} & Z_{n\text{stepEC}} & Z_{n\text{stepE}} & Z_{n\text{stepC}} & Z_{n\text{logEC}} & Z_{n\text{logE}} & Z_{n\text{logC}} \end{bmatrix}. \tag{B.3}$$

Let $\boldsymbol{\beta_{TIME}} =$

$$\boldsymbol{\beta_{TIME}} = (\beta_{\text{linEC}}, \beta_{\text{linE}}, \beta_{\text{linC}}, \beta_{\text{stepEC}}, \beta_{\text{stepE}}, \beta_{\text{stepC}}, \beta_{\text{logEC}}, \beta_{\text{logE}}, \beta_{\text{logC}})^T. \tag{B.4}$$

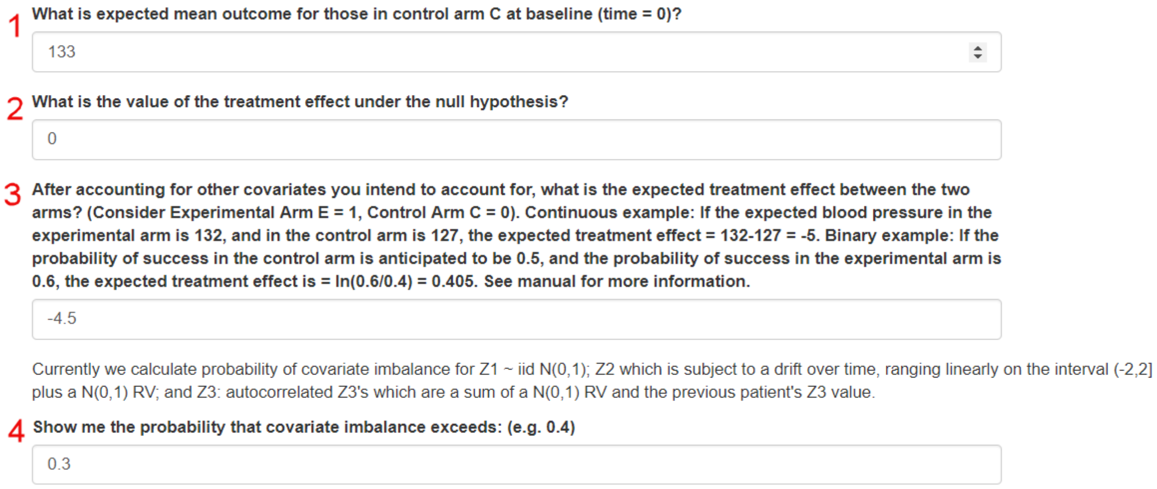
For continuous responses, the true response is defined as:

$$E(Y|j,h) = \boldsymbol{X\beta} + \boldsymbol{Z\beta_{TIME}}. \tag{B.5}$$

The naive and adjusted analyses are defined as:

$$\textbf{Naive}\text{: } E(Y|j,h) = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{B.6}$$

$$\textbf{Adjusted}\text{: } E(Y|j,h) = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\beta_{TIME}}}. \tag{B.7}$$



1 What is expected mean outcome for those in control arm C at baseline (time = 0)?

133

2 What is the value of the treatment effect under the null hypothesis?

0

3 After accounting for other covariates you intend to account for, what is the expected treatment effect between the two arms? (Consider Experimental Arm E = 1, Control Arm C = 0). Continuous example: If the expected blood pressure in the experimental arm is 132, and in the control arm is 127, the expected treatment effect = 132-127 = -5. Binary example: If the probability of success in the control arm is anticipated to be 0.5, and the probability of success in the experimental arm is 0.6, the expected treatment effect is = ln(0.6/0.4) = 0.405. See manual for more information.

-4.5

Currently we calculate probability of covariate imbalance for Z1 ~ iid N(0,1); Z2 which is subject to a drift over time, ranging linearly on the interval (-2,2] plus a N(0,1) RV; and Z3: autocorrelated Z3's which are a sum of a N(0,1) RV and the previous patient's Z3 value.

4 Show me the probability that covariate imbalance exceeds: (e.g. 0.4)

0.3

Figure B.6

1. What is the expected mean outcome for those in control arm C at baseline (time = 0)?

If you believe that the responses in both treatment arms are free from time trends, your response will be the same as in Question 3 shown in Figure B.4 (What is the expected mean outcome for those in control arm C?). If there is a potential time trend, input the expected outcome of a subject in the control arm at the very start of the trial.

2. What is the value of the treatment effect under the null hypothesis?
Typically, this is 0.
$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

3. After accounting for other covariates you intend to account for, what is the expected treatment effect between the two arms?...
This is asking what you expect β_1 to equal; how much you expect the *experimental treatment* to alter the response for the experimental arm relative to the control arm. Here are two examples:
Binary:
$H_0 : p_E - p_C = 0$
$H_1 : p_E - p_C \neq 0$. Specifically, $p_E = 0.7$, $p_C = 0.4$. Then, the probability of success for a subject in the control arm (C) is:

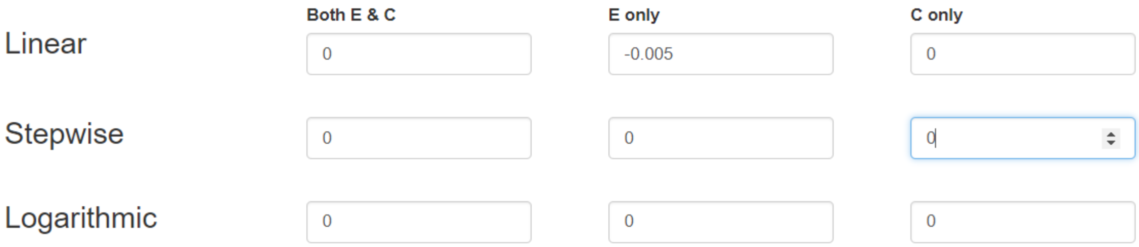
$$\boldsymbol{p} = \text{Probability}(Y = 1|j, h = 1) = 0.4 = \frac{exp(\beta_0 + \beta_1 \times 0)}{1 + exp(\beta_0 + \beta_1 \times 0)}.$$

Then, $\beta_0 = ln(\frac{0.4}{0.6}) = -0.405$. The probability of success for a subject in the experimental arm (E) is:

$$\boldsymbol{p} = \text{Probability}(Y = 1|j, h = 1) = 0.7 = \frac{exp(ln(\frac{0.4}{0.6}) + \beta_1 \times 1)}{1 + exp(ln(\frac{0.4}{0.6}) + \beta_1 \times 1)}.$$

Then, for this question, input $\beta_1 = ln((\frac{0.7}{0.3})/\frac{0.4}{0.6}) = 1.253$.
Continuous:
$H_0 : \mu_E - \mu_C = 0$
$H)1 : \mu_E - \mu_C \neq 0$. Specifically, $\mu_E = 127$, $\mu_C = 132$. Then, the expected response for a subject in the control arm (C) is: $E(Y) = 132 = \beta_0 + \beta_1 \times 0$, resulting in $\beta_0 = 132$. The expected response for a subject in the experimental arm (E) is: $E(Y) = 127 = 132 + \beta_1 \times 1$, resulting in the response for this question being $\beta_1 = -5$.

4. Show me the probability that covariate imbalance exceeds: (e.g. 0.4)
The program calculates this by observing the proportion of simulated trials where the covariate imbalance as defined previously exceeds the value you specify.



	Both E & C	E only	C only
Linear	0	-0.005	0
Stepwise	0	0	0
Logarithmic	0	0	0

Figure B.7

If time trends are not a concern during the design phase of your study, input 0 for all 9 boxes in this 3X3 grid. If the 3X3 grid is filled with all 0's, this is the last question of Step 1. You can click on "Show me my Design!" to see summaries of characteristics of the designs under evaluation.

Otherwise, input the value of $\boldsymbol{\beta}_{\text{TIME}}$, which is the coefficient for a time trend covariate $\boldsymbol{Z}_{\text{TIME}}$. The definition of Z_{TIME} depends on whether you believe patient enrollment is uniform (the rate of patient enrollment is consistent throughout the trial), or whether you would like to simulate varying enrollment patterns throughout the trial.

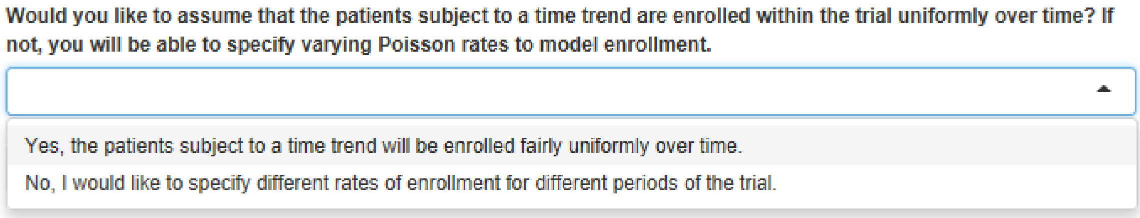
If enrollment into the trial is uniform,

$Z_{j\text{linEC}} = (j-1)$, $Z_{j\text{linE}} = (j_E - 1)$, $Z_{j\text{linC}} = (j_C - 1)$,
$Z_{j\text{logEC}} = log(j)$, $Z_{j\text{logE}} = log(j_E)$, $Z_{j\text{logC}} = log(j_C)$, and
$Z_{j\text{stepEC}} = \mathbb{1}(j > t)$, $Z_{j\text{stepE}} = \mathbb{1}(j_E > t)$, $Z_{j\text{stepC}} = \mathbb{1}(j_C > t)$.

For example, with uniform enrollment is a linear time trend, β_{TIME} is the expected change in response for each sequential enrolled subject affected by the trend. For example, if $n = 128$ and you expect a linear trend in both E & C, a value of $\beta_{\text{TIME}} = 1/127$ in the first element of the 3X3 grid would indicate $E(Y) = \mu_E T_j + \mu_C(1 - T_j) + (1/127)(j-1)$, indicating that the 128th enrolled subject has an expected response 1 unit higher than that of the 1st enrolled subject, and the expected responses for the subjects between them increase with a linear pattern. With a log trend affecting the experimental arm only, a value of $\beta_{\text{TIME}} = 1/log(64)$ in the "E only" column of the "logarithmic" row would indicate that the 64th enrolled subject of the experimental arm would have an expected response 1 unit higher than that of the 1st enrolled subject, and the expected responses for the subjects between them increase with a logarithmic pattern. With a stepwise trend affecting the control arm only, a value of $\beta_{\text{TIME}} = 1$ and $t = 33$ would indicate that after 33 subjects have enrolled into the control arm, the expected response shifts upwards by 1 unit.

On the other hand, if enrollment into the trial is not uniform, measurement time of the response plays a role in the definition of Z. Let $G \in [EC, E, C]$ represent whether the time trend affects both the experimental and control groups, the experimental group only, or the control group only.

$Z_{j\text{linG}} = m_j$,
$Z_{j\text{logG}} = \log(m_j)$, and
$Z_{j\text{stepG}} = \mathbb{1}(m_j \geq step \times M)$.



Would you like to assume that the patients subject to a time trend are enrolled within the trial uniformly over time? If not, you will be able to specify varying Poisson rates to model enrollment.

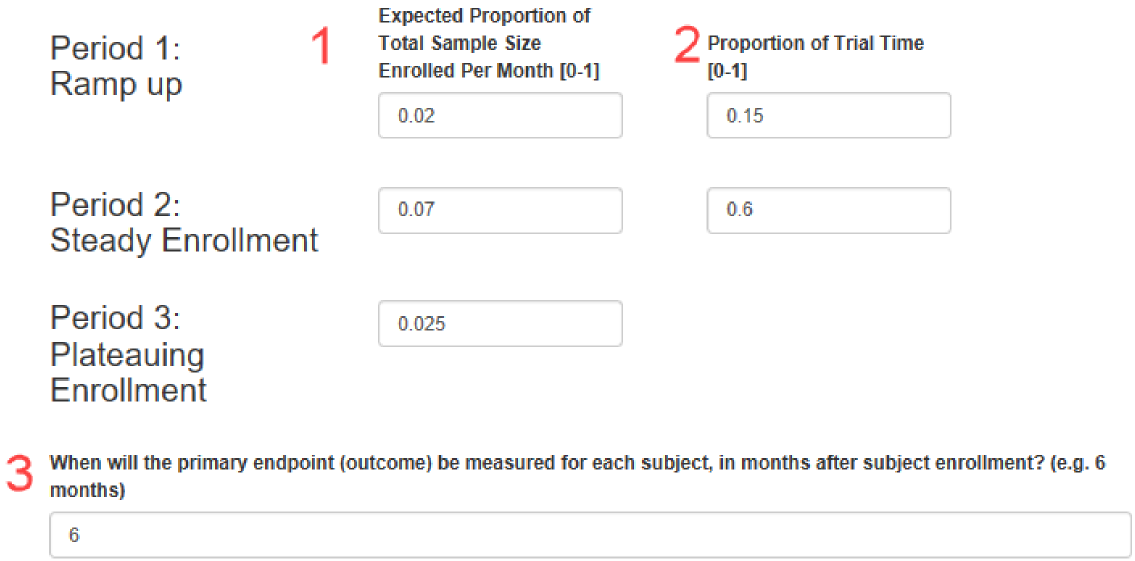
▲
Yes, the patients subject to a time trend will be enrolled fairly uniformly over time.
No, I would like to specify different rates of enrollment for different periods of the trial.

Figure B.8

1. Would you like to assume that the patients subject to a time trend are enrolled within the trial uniformly over time? If not, you will be able to specify varying Poisson rates to model enrollment.
Select "Yes, the patients subject to a time trend will be enrolled fairly uniformly over time." if you expect subjects to be recruited and enrolled at a consistent rate from the beginning of the trial until the required sample size is met. If you select "Yes", this will be the last question of Step 1, and you can click on "Show me my Design!" to see summaries of characteristics of the designs under evaluation.

Otherwise, select "No, I would like to specify different rates of enrollment for different periods of the trial."

1. Expected Proportion of Total Sample Size Enrolled Per Month [0-1].
Input the proportion of total sample size you expect to enroll per month during the Ramp Up, Steady



	Expected Proportion of Total Sample Size Enrolled Per Month [0-1]	Proportion of Trial Time [0-1]
Period 1: Ramp up	0.02	0.15
Period 2: Steady Enrollment	0.07	0.6
Period 3: Plateauing Enrollment	0.025	

When will the primary endpoint (outcome) be measured for each subject, in months after subject enrollment? (e.g. 6 months)

6

Figure B.9

Enrollment, and Plateauing Enrollment Periods. If you only have 2 Periods, input 0 for Period 3.

In this example, Period 1: Ramp Up has a value of 0.02 input. This means that 2% of the total sample size n is expected to be enrolled into the trial each month during Period 1. Similarly, 7% and 2.5% of n is expected to be enrolled per month during Period 2: Steady Enrollment and Period 3: Plateauing Enrollment, respectively.

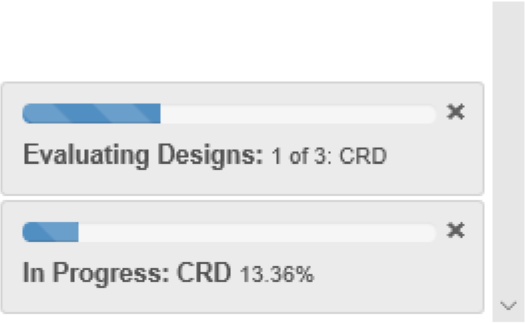
2. Proportion of Trial Time [0-1].
In this example, Period 1: Ramp Up has a value of 0.15 input here. This means that 15% of the total recruitment or trial time is enrolling subjects at the Ramp Up Rate ($0.02n$/month in this example). Similarly, the value of 0.60 input for Steady Enrollment indicates that 60% of te total recruitment or trial time is enrolling subjects at the Ramp Up Rate ($0.07n$/month in this example). Lastly, the proportion of time at the Plateauing Enrollment rate is 1-proportion of trial time in ramp up period - proportion of trial time in steady enrollment period = 1-0.15-0.6=0.25 in this example. This means that the last 25% of the trial time enrolls at the Plateauing Enrollment rate of $0.025n$/month.

3. When will the primary endpoint (outcome) be measured for each subject, in months after subject enrollment?
Input here when the response for each subject will be measured, in months after enrollment.

While questions 1 and 2 helped the program simulate enrollment times, the important variable in assessing the impact of time trends are Z_{TIME}, which is the measurement time. For example, if the first patient ($j = 1$) is enrolled at time = 0, their response is measured at 6 months. If the 40th patient ($j = 40$) is enrolled at month 6, then their response is measured at 12 months.

After answering this, you will have completed the questions of Step 1. Click on “Show me my Design!” to see summaries of characteristics of the designs under evaluation.

After clicking “Show me my Design”, two progress bars will indicate where in the simulation the program is.



Evaluating Designs: 1 of 3: CRD

In Progress: CRD 13.36%

Figure B.10

Interpretation of Results:

See Evaluated Characteristics of Designs for more background regarding the characteristics summarized in this section.

Simulated Recruitment

This output is displayed only if you indicated that patients will not be recruited at a consistent rate throughout the recruitment period.

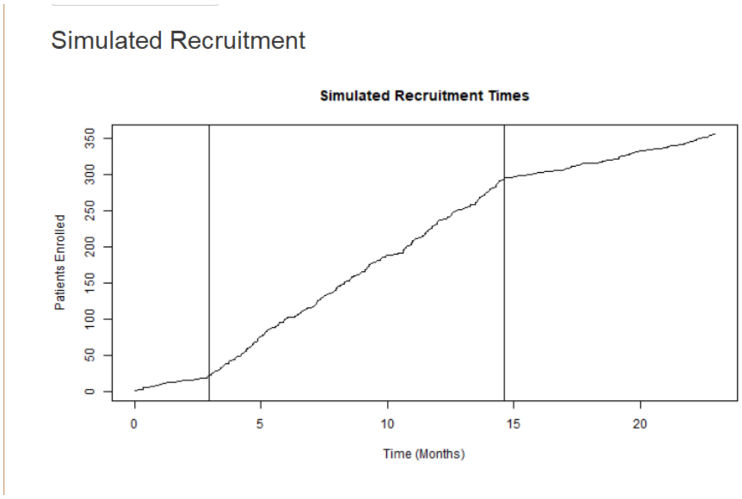


Figure B.11

Characteristics of Treatment Allocation

- patientsinE_mean: the number of patients placed in the experimental arm E averaged across all simulated trials (i.e. number of iterations).
- patientsinE_sd: the standard deviation of the number of patients placed in the experimental arm E across all simulated trials.

- proportioninE: The proportion of subjects placed in the experimental arm (n_E/n) averaged across all simulated trials.
- mintrtimb: minimum treatment imbalance ($n_E - n_C$) witnessed across all simulated trials.
- meantrtimb: the average treatment imbalance ($n_E - n_C$) across all simulated trials.
- mediantrtimb: the median treatment imbalance ($n_E - n_C$) across all simulated trials.
- maxtrtimb: the maximum treatment imbalance ($n_E - n_C$) witnessed across all simulated trials.

Characteristics of Outcome

- meanuE_h0: the average response in the experimental arm in simulated trials where the null hypothesis is true.
- meanuE_h1: the average response in the experimental arm in simulated trials where the alternative hypothesis is true.
- meanuC_h0: the average response in the control arm in simulated trials where the null hypothesis is true.
- meanuC_h1: the average response in the control arm in simulated trials where the alternative hypothesis is true.

Summary of Failures

- number_failures_h1_mean: the average number of failures in the experimental arm across simulated trials where the alternative hypothesis is true.
- number_failures_h1_sd: the standard deviation of the number of failures in the experimental arm across simulated trials where the alternative hypothesis is true.

Simple T-test Results (Estimating Treatment Effect Only)

- esttrtdiff_h1_mean: the difference in means $\bar{Y}_E - \bar{Y}_C$ of the two treatment groups averaged across all simulated trials.
- esttrtdiff_h1_sd: the standard deviation of the difference in means $\mathrm{sd}(\bar{Y}_E - \bar{Y}_C)$ of the two treatment groups across all simulated trials.
- esttrtdiffMSE_h1: the mean squared error of the difference in means of the two treatment groups across all simulated trials. The mean squared error is defined as $\mathrm{bias}^2 + \mathrm{variance}$: $((\bar{Y}_E - \bar{Y}_C) - (\mu_E - \mu_C))^2 + sd(\bar{Y}_E - \bar{Y}_C^2)$.
- typeIerror_h0: the proportion of times the null hypothesis is rejected in all simulated trials, where the null hypothesis is true.
- power_h1: the proportion of times the null hypothesis is rejected in all simulated trials, where the alternative hypothesis is true.

Summary of Imbalance of 3 Different Covariates

See the section on Covariate Imbalance.

- probcovimb1_h0: $\mathrm{Probability}(\left|\overline{Z1}_E - \overline{Z1}_C\right| > \epsilon) = \frac{\sum_{i=1}^{\mathrm{iter}} \mathbb{1}(\left|\overline{Z1}_{E_i} - \overline{Z1}_{C_i}\right| > \epsilon)}{\mathrm{iter}}$, where each iteration simulates a study where the null hypothesis is true.
- probcovimb2_h0: $\mathrm{Probability}(\left|\overline{Z2}_E - \overline{Z2}_C\right| > \epsilon) = \frac{\sum_{i=1}^{\mathrm{iter}} \mathbb{1}(\left|\overline{Z2}_{E_i} - \overline{Z2}_{C_i}\right| > \epsilon)}{\mathrm{iter}}$, where each iteration simulates a study where the null hypothesis is true.

- probcovimb3_h0: Probability($|\overline{Z3}_E - \overline{Z3}_C| > \epsilon$) $= \frac{\sum_{i=1}^{\text{iter}} \mathbb{1}(|\overline{Z3}_{E_i} - \overline{Z3}_{C_i}|>\epsilon)}{\text{iter}}$, where each iteration simulates a study where the null hypothesis is true.
- probcovimb1_h1, probcovimb2_h1,
- probcovimb3_h1: same as above, except each iteration simulates a study where the alternative hypothesis is true.

Summary of Accidental Bias Factor

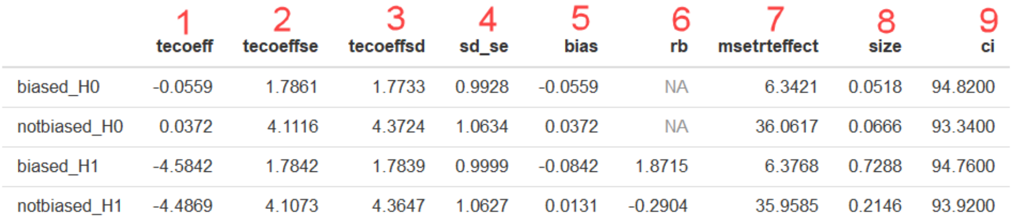
- minaccbias: the minimum accidental bias factor estimate across all simulated trials.
- meanaccbias: the average accidental bias factor estimate across all simulated trials.
- maxaccbias: the maximum accidental bias factor estimate across all simulated trials.

Summary of Selection Bias

- minselbias_h1: the minimum selection bias as measured by the predictability ρ_{pred} of a randomization sequence across all simulated trials.
- meanselbias_h1: the average selection bias as measured by the predictability ρ_{pred} of a randomization sequence across all simulated trials.
- maxselbias_h1: the maximum selection bias as measured by the predictability ρ_{pred} of a randomization sequence across all simulated trials.

Summary of Chronological Bias

The summary for a single design looks like the table shown in Figure B.12.



	1	2	3	4	5	6	7	8	9
	tecoeff	tecoeffse	tecoeffsd	sd_se	bias	rb	msetrteffect	size	ci
biased_H0	-0.0559	1.7861	1.7733	0.9928	-0.0559	NA	6.3421	0.0518	94.8200
notbiased_H0	0.0372	4.1116	4.3724	1.0634	0.0372	NA	36.0617	0.0666	93.3400
biased_H1	-4.5842	1.7842	1.7839	0.9999	-0.0842	1.8715	6.3768	0.7288	94.7600
notbiased_H1	-4.4869	4.1073	4.3647	1.0627	0.0131	-0.2904	35.9585	0.2146	93.9200

Figure B.12

The tables are shown in the order a shown in the “Designs selected are:” at the top of this page (Step 1). Here, the designs selected were CRD, PBD, and DBCD.RSIHR, so 3 tables like the one shown in Figure B.12 would appear.

Remember the true model is as shown in Equation B.5.
The four rows in the table are:

- biased_H0: naive analysis (Equation B.6) when the null hypothesis is true
- notbiased_H0: adjusted analysis (Equation B.7) when the null hypothesis is true
- biased_H1: naive analysis (Equation B.6) when the alternative hypothesis is true

- notbiased_H1: adjusted analysis (Equation B.7) when the alternative hypothesis is true

The columns are:

- **1** tecoeff: Average estimate of the treatment effect $\hat{\beta}_1$ across all simulated trials
- **2** tecoeffse: Standard error of the treatment effect estimate $\hat{\beta}_1$ across all simulated trials
- **3** tecoeffsd: Standard deviation of the treatment effect estimate $\hat{\beta}_1$ across all simulated trials
- **4** sd_se: Ratio of **2/3**
- **5** bias: $\hat{\beta}_1 - \beta$
- **6** relative bias: $\frac{\hat{\beta}_1 - \beta}{\beta} \times 100$
- **7** msetrteffect: mean squared error of the treatment effect estimate: $\mathbf{5}^2 + \mathbf{3}^2$
- **8** size: empirical test size = proportion of times H_0 is rejected. For H_0 this is Type I error. For H_1 this is power.
- **9** ci: empirical coverage = proportion of times the 95% confidence interval for the treatment effect includes the true value β_1.

Click "Next" to proceed to Step 2. Individual Desirability Functions.

Step 2.

There are two types of desirability scores: *individual* and *overall*. In this step, *individual* desirability functions are defined. The characteristics under evaluation are given an "individual desirability score" between 0 and 1 using these individual desirability functions. A desirability score of 0 indicates "completely unacceptable", while a desirability score of 1 represents "the ultimate in satisfaction and quality, where an improvement beyond this point would have no additional meaningful value". A table with guided interpretations of desirability scores is provided in Table B.1. Later, these scores are combined into a single overall score for each design.
An overall rating system using the overall desirability was provided as shown in Table B.1.

Desirability Score	**Interpretation**
1.00	The ultimate in satisfaction and quality, where an improvement beyond this point would have no additional meaningful value
[0.80, 1.00)	Acceptable and excellent (represents unusual quality or performance well beyond anything commercially available)
[0.63, 0.80)	Acceptable and good (represents an improvement over the best commercial quality)
[0.40, 0.63)	Acceptable but poor (quality is acceptable to the specification limits but improvement is desired)
[0.30, 0.40)	Borderline (if specification exists, then some of the product quality lies exactly on the specification maximum or minimum)
(0.00, 0.30)	Unacceptable (materials of this quality would lead to failure)
0.00	Completely unacceptable

Table B.1: Interpretations of desirability scores.

This website allows you to construct an individual desirability function in two ways: using a scale method, or using a mapping method. An example of each of these methods is shown in Figure B.13.

Scale Method:
Individual desirability functions are shaped my scale parameters. For the characteristic under evaluation, there are

- smaller-the-better: the smaller the value of the characteristic, the better (e.g. Type I error)
- larger-the-better: the larger the value of the characteristic, the better (e.g. power)
- nominal-the-better: a target value is best, while values less than or greater than this target indicate diminishing of quality (e.g. treatment group size imbalance with a target value of 0).

Let y_i be the value of a characteristic. Let U and L be the upper- and lower- bounds of the characteristic's acceptable values, respectively. Let T is the target value of a nominal-the-better type response. Shape parameters r, r_1, and $r_2 > 0$ are selected by the user. These parameters are used to calculate d_i, the individual desirability score of the value of y_i.

For Larger the Better (LTB):

$$d_i = \begin{cases} 0 & \text{for } y_i \leq L \\ (\frac{y_i - L}{U - L})^r & \text{for } L < y_i < U \\ 1 & \text{for } y_i \geq U \end{cases} \tag{B.8}$$

For Smaller the Better (STB):

$$d_i = \begin{cases} 1 & \text{for } y_i \leq L \\ (\frac{U - y_i}{U - L})^r & \text{for } L < y_i < U \\ 0 & \text{for } y_i \geq U \end{cases} \tag{B.9}$$

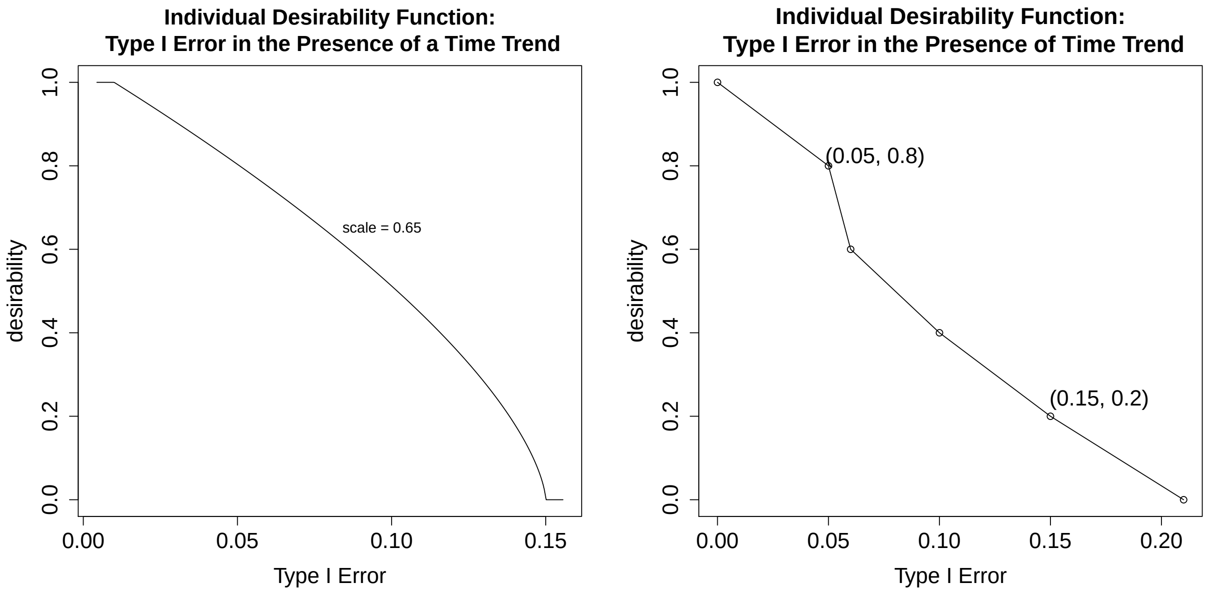
For Nominal The Better (NTB):

$$d_i = \begin{cases} (\frac{y_i - L}{T - L})^{r_1} & \text{for } L \leq y_i \leq T \\ (\frac{U - y_i}{U - T})^{r_2} & \text{for } T < y_i \leq U \\ 0 & \text{for } y_i < L \text{ or } y_i > U \end{cases} \tag{B.10}$$

Mapping Method:
The mapping method maps specific individual desirability scores to specific values of a component. Individual points of the mapping are connected linearly so an individual desirability curve is created.

In Figure B.13a, the scaling method is used with Equation B.9 and scale parameter $r = 0.65$, $L = 0.01$, and $U = 0.15$. This means that any Type I errors below 0.01 will receive an individual desirability score of 1, and any Type I errors above 0.15 will receive an individual desirability score of 0. It can be seen that a Type I error of 0.05 yields an individual desirability of approximately 0.8, highly desirable. In Figure B.13b, a mapping of different Type I errors to different desirability scores yielded this plot. It can be seen that a Type I error yields a desirability score of 0.8 (highly desirable), and a Type I error of 0.15 yields a desirability score of 0.2 (unacceptable). The mapping that results in this function is Type I errors of (0, 0.05, 0.06, 0.1, 0.15, 0.21) for individual desirability scores of (1, 0.8, 0.6, 0.4, 0.2, 0), respectively.

The program asks if you would like to use a scale parameter for each assessed component (scale method),

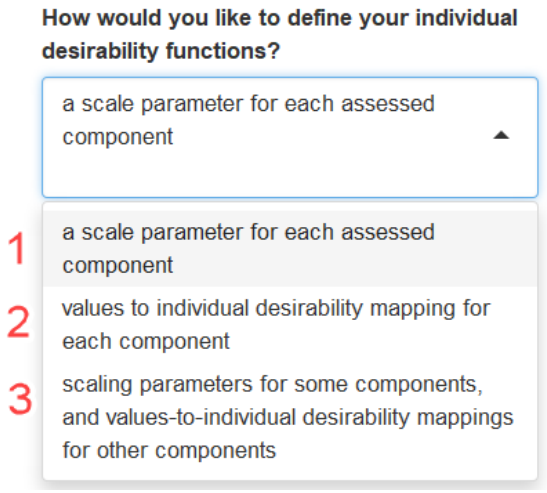


(a) Scale Method

(b) Mapping Method

Figure B.13: Individual Desirability Function: Type I Error.

values to individual desirability mapping for each component (mapping method), or scaling parameters for some components, and values-to-individual desirability mappings for other components.


How would you like to define your individual desirability functions?

a scale parameter for each assessed component ▲

1 a scale parameter for each assessed component

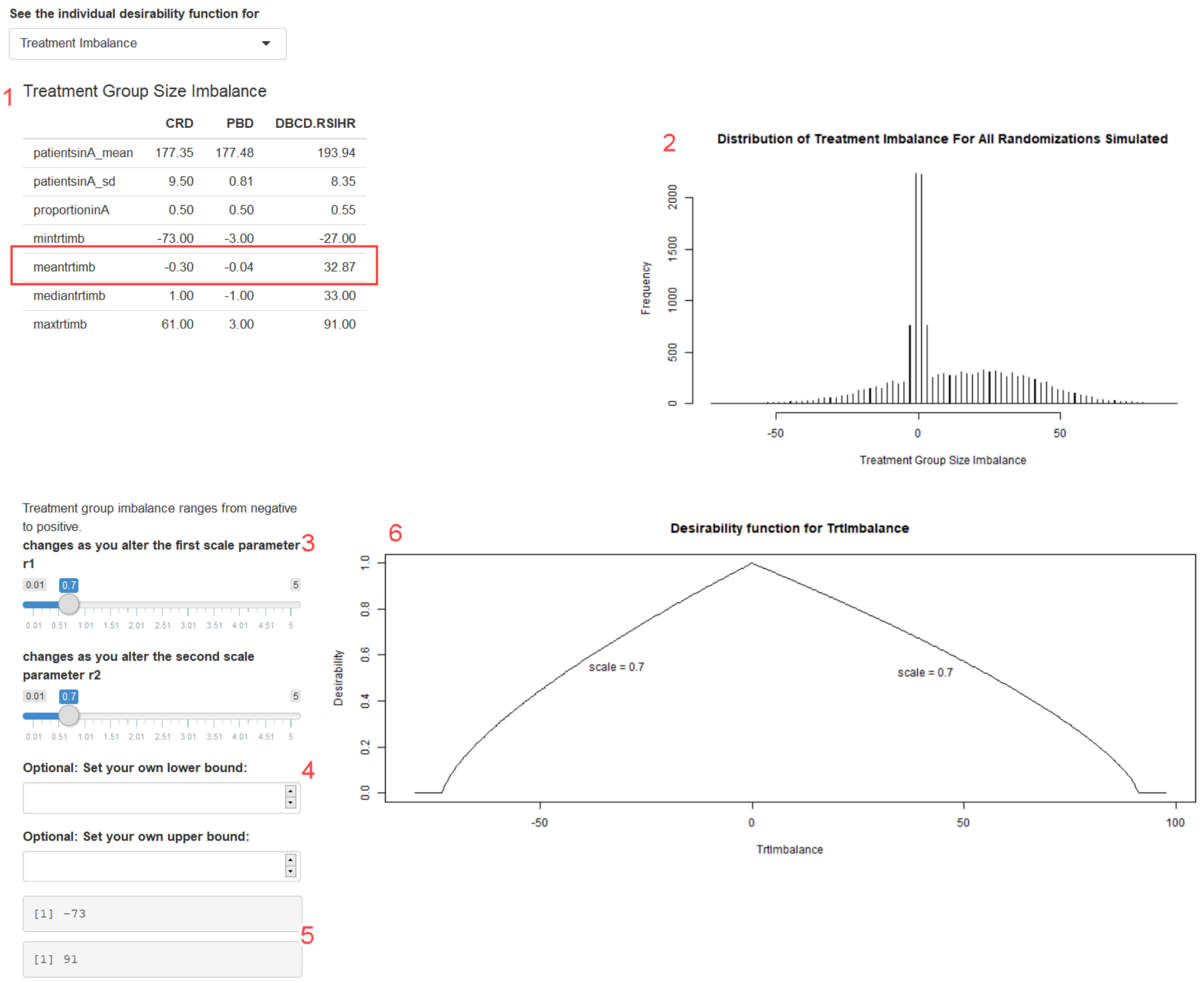
2 values to individual desirability mapping for each component

3 scaling parameters for some components, and values-to-individual desirability mappings for other components

Figure B.14

If selecting 1 (a scale parameter for each assessed component), use the drop-down menu to see the individual desirability functions for each component one at a time. For example, selecting "Treatment Imbalance" in our example leads to:

1 This is the same summary table as shown at the end of Step 1. It is shown again to help users understand the range of values each characteristic for each design takes. This may help one's decision-making in deciding what type of individual desirability function/curve best reflects their own judgment. Although summary statistics are shown, the mean is used to calculate the individual desirability score.



See the individual desirability function for

Treatment Imbalance

1 Treatment Group Size Imbalance

	CRD	PBD	DBCD.RSIHR
patientsinA_mean	177.35	177.48	193.94
patientsinA_sd	9.50	0.81	8.35
proportioninA	0.50	0.50	0.55
mintrtimb	-73.00	-3.00	-27.00
meantrtimb	-0.30	-0.04	32.87
mediantrtimb	1.00	-1.00	33.00
maxtrtimb	61.00	3.00	91.00

Treatment group imbalance ranges from negative to positive.

changes as you alter the first scale parameter r1

changes as you alter the second scale parameter r2

Optional: Set your own lower bound:

Optional: Set your own upper bound:

```
[1] -73
```

```
[1] 91
```

Figure B.15

2 This is a histogram showing the distribution of treatment imbalance $n_E - n_C$ for all randomizations simulated.

3 Since the mean treatment imbalance ranges from negative to positive values in this example, the program decides that this characteristic is “nominal-the-better”, with a target value of 0. Here, set the scale parameters r1 and r2 by clicking on the slider or dragging the dot to an appropriate point on the slider. The graph in **6** will automatically adjust so you can see if the curve accurately reflects your preferences. For example, if you prefer $n_E > n_C$, you may want to penalize positive values of treatment imbalance less than negative values of treatment imbalance.

4 Optional: Set your lower (or upper) bounds. Input your lower limit L and upper limit U here for the desirability function. If you leave these blank, the program will automatically decide for you. Here, the default values for L and U are the minimum treatment imbalance and maximum treatment imbalance values witnessed in the simulations. See **5**.

5 This displays the L and U the program will use. If you have input values in **4**, they will be displayed here.

6 This is the individual desirability function, plotting individual desirability score versus the characteristic under evaluation.

If selecting **2** from Figure B.14, you will be asked to inform the program which values deserve scores of 1, 0.8, 0.6, 0.4, 0.2, and 0. For STB and LTB characteristics, one value will map to each of these scores. For NTB characteristics, two values will map to each of these scores (one less than and one greater than the target value T). The individual desirability function connects lines between these values. For STB (LTB) characteristics, values smaller (larger) than the value mapped to an individual desirability score of 1 will automatically receive a score of 1, and values larger (smaller) than the value mapped to a score of 0 will receive a score of 0.

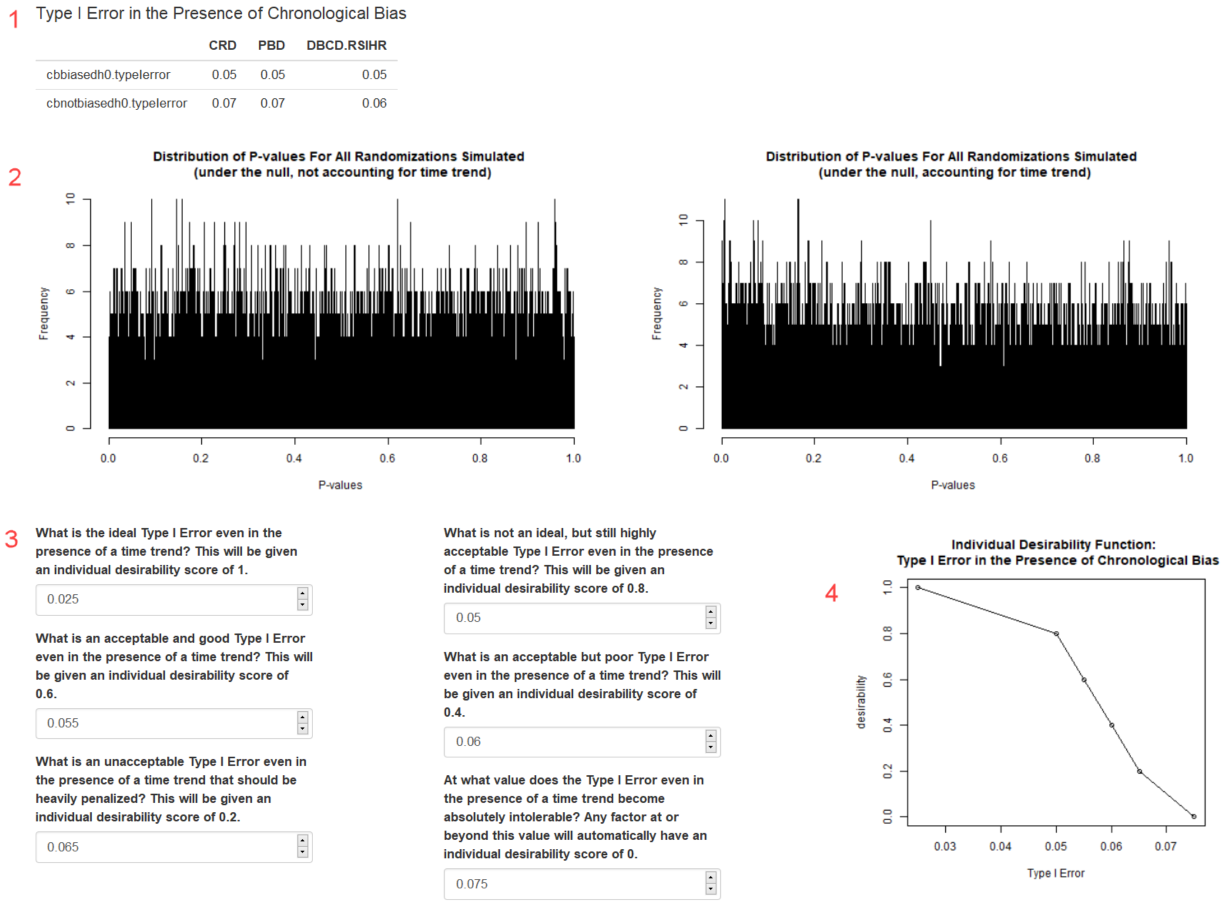
1 This is the same summary table as shown at the end of Step 1. It is shown again to help users understand the range of values each characteristic for each design takes. This may help one’s decision-making in deciding what type of individual desirability function/curve best reflects their own judgment. Although summary statistics are shown, the mean is used to calculate the individual desirability score.

2 This is a histogram showing the distribution of p-values of the treatment effect estimate in Equation B.6-B.7 for all randomizations simulated.

3 This is the mapping process. The program asks the user to input which values of the characteristic should be assigned individual desirability scores of 1, 0.8, 0.6, 0.4, 0.2, and 0.

4 This is the individual desirability function, plotting individual desirability score versus the characteristic under evaluation. Note how the desirability scores between the mapped values are connected linearly.

If selecting **3** from Figure B.14, you will be asked to inform the program which characteristics you would like to build individual desirability methods using the scale method (check the appropriate checkboxes).



Type I Error in the Presence of Chronological Bias

	CRD	PBD	DBCD.RSIHR
cbbiasedh0.typeIerror	0.05	0.05	0.05
cbnotbiasedh0.typeIerror	0.07	0.07	0.06

2
Distribution of P-values For All Randomizations Simulated
(under the null, not accounting for time trend)
Frequency
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
P-values
Distribution of P-values For All Randomizations Simulated
(under the null, accounting for time trend)
Frequency
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
P-values
3

What is the ideal Type I Error even in the presence of a time trend? This will be given an individual desirability score of 1.

0.025

What is an acceptable and good Type I Error even in the presence of a time trend? This will be given an individual desirability score of 0.6.

0.055

What is an unacceptable Type I Error even in the presence of a time trend that should be heavily penalized? This will be given an individual desirability score of 0.2.

0.065

What is not an ideal, but still highly acceptable Type I Error even in the presence of a time trend? This will be given an individual desirability score of 0.8.

0.05

What is an acceptable but poor Type I Error even in the presence of a time trend? This will be given an individual desirability score of 0.4.

0.06

At what value does the Type I Error even in the presence of a time trend become absolutely intolerable? Any factor at or beyond this value will automatically have an individual desirability score of 0.

0.075

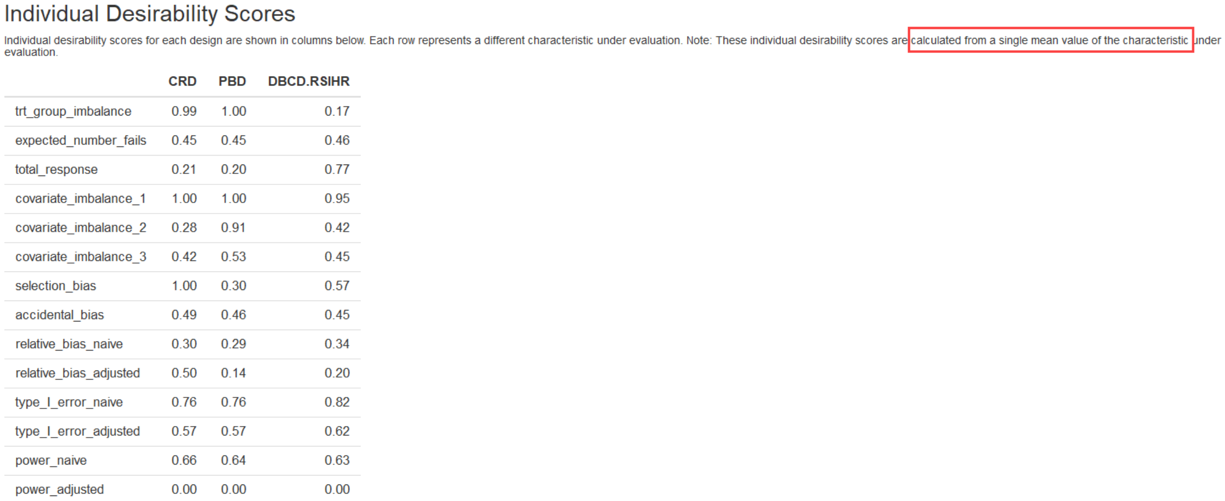
4
Individual Desirability Function:
Type I Error in the Presence of Chronological Bias
desirability
0.0
0.2
0.4
0.6
0.8
1.0
0.03
0.04
0.05
0.06
0.07
Type I Error

Figure B.16

You will then be asked to define these characteristics' individual desirability functions first, followed by the definitions of characteristics whose individual desirability functions are defined by the mapping method. See instructions from examples highlighted in Figures B.15-B.16 to understand how to define individual desirability functions for both the scale and mapping methods.

After defining the individual desirability functions, click on "Calculate Individual Desirabilities Using These Scale Parameters", "Calculate Individual Desirabilities Using These Mappings", or "Calculate Individual Desirabilities Using These Scale Parameters & Mappings".
If you receive a red error message, check to ensure you have informed the program to define individual desirability functions for each of the characteristics. Commonly, the values for the mapping method for "Total Response" characteristic are forgotten. The program is unable to autofill these for you at this time.

Individual desirability scores are calculated using two different ways. The first is shown here:



Individual Desirability Scores

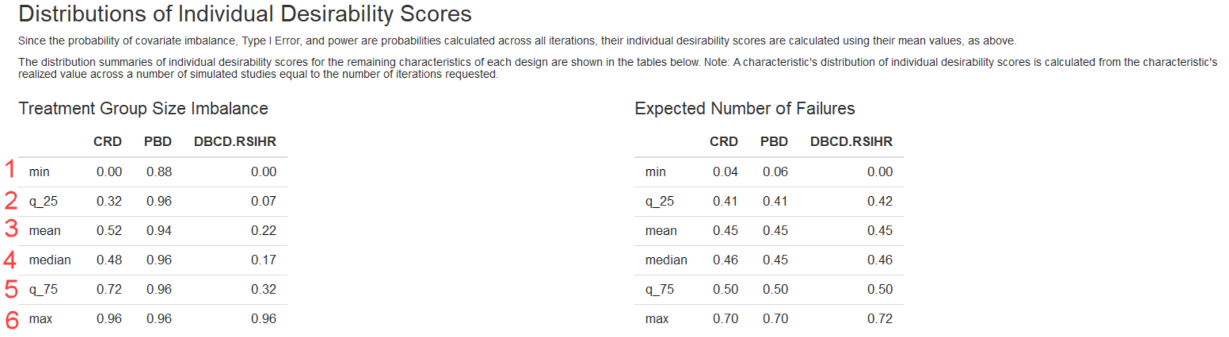
Individual desirability scores for each design are shown in columns below. Each row represents a different characteristic under evaluation. Note: These individual desirability scores are calculated from a single mean value of the characteristic under evaluation.

	CRD	PBD	DBCD.RSIHR
trt_group_imbalance	0.99	1.00	0.17
expected_number_fails	0.45	0.45	0.46
total_response	0.21	0.20	0.77
covariate_imbalance_1	1.00	1.00	0.95
covariate_imbalance_2	0.28	0.91	0.42
covariate_imbalance_3	0.42	0.53	0.45
selection_bias	1.00	0.30	0.57
accidental_bias	0.49	0.46	0.45
relative_bias_naive	0.30	0.29	0.34
relative_bias_adjusted	0.50	0.14	0.20
type_I_error_naive	0.76	0.76	0.82
type_I_error_adjusted	0.57	0.57	0.62
power_naive	0.66	0.64	0.63
power_adjusted	0.00	0.00	0.00

Figure B.17

These individual desirability scores are calculated from a *single* value (the mean) of each characteristic for each design. In other words, the mean value for each characteristic is used to calculate a single individual desirability score.

The second is shown here:



Distributions of Individual Desirability Scores

Since the probability of covariate imbalance, Type I Error, and power are probabilities calculated across all iterations, their individual desirability scores are calculated using their mean values, as above.

The distribution summaries of individual desirability scores for the remaining characteristics of each design are shown in the tables below. Note: A characteristic's distribution of individual desirability scores is calculated from the characteristic's realized value across a number of simulated studies equal to the number of iterations requested.

Treatment Group Size Imbalance

		CRD	PBD	DBCD.RSIHR
1	min	0.00	0.88	0.00
2	q_25	0.32	0.96	0.07
3	mean	0.52	0.94	0.22
4	median	0.48	0.96	0.17
5	q_75	0.72	0.96	0.32
6	max	0.96	0.96	0.96

Expected Number of Failures

	CRD	PBD	DBCD.RSIHR
min	0.04	0.06	0.00
q_25	0.41	0.41	0.42
mean	0.45	0.45	0.45
median	0.46	0.45	0.46
q_75	0.50	0.50	0.50
max	0.70	0.70	0.72

Figure B.18

This displays a *distribution* of individual desirability scores, which is calculated using the *distribution* of values of the characteristic across each iterations. This means that for *each* characteristic under evaluation, the number of individual desirability scores calculated is equal to the number of iterations you have informed the program to simulate. The summary statistics of the distribution of individual desirability scores of each characteristic is shown, with rows representing:

- **1** min: the minimum
- **2** q_25: the 25th percentile
- **3** mean: the average
- **4** median: the median
- **5** q_75: the 75th percentile
- **6** max: the maximum

individual desirability score calculated from all values of the characteristic (treatment group size imbalance, in this case) realized in *iter* simulated trials.

Note that certain characteristics are already averaged across all iterations (Total Response, probability of covariate imbalance, Type I error, Power), so they are excluded from this method.

Click on **Next** to proceed to Step 3.

Step 3.

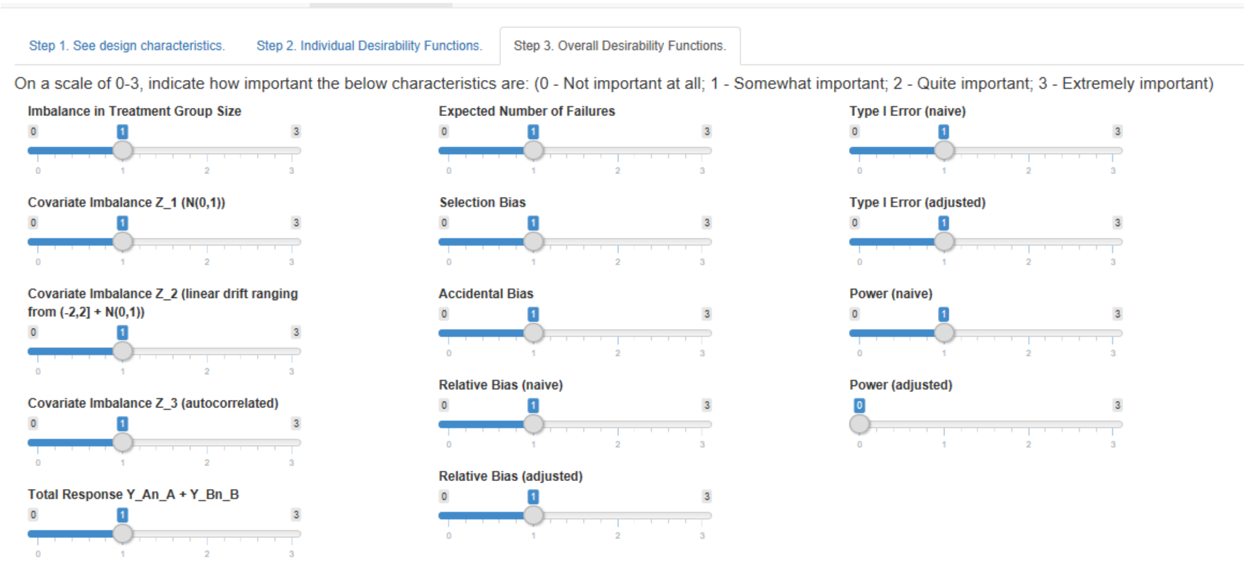
In this step, *overall* desirability functions are calculated. The overall desirability score is calculated as a weighted geometric mean of the individual desirability scores:

$$D = (d_1^{w_1} d_2^{w_2} \ldots d_m^{w_m})^{\frac{1}{\sum_{i=1}^{m} w_i}}, \tag{B.11}$$

where w_i is the weight assigned to each individual desirability d_i, with larger weights indicating more importance.

In Step 3, the program asks you to rate on a scale between 0 to 3 the relative importance of each characteristic under evaluation. The resulting overall desirability scores will display below. You can slide your responses and see how the overall desirability scores change in real time.

First, you will see a table labeled "**Individual Desirability Scores** Calculated from mean attribute values":



Step 1. See design characteristics.
Step 2. Individual Desirability Functions.
Step 3. Overall Desirability Functions.

On a scale of 0-3, indicate how important the below characteristics are: (0 - Not important at all; 1 - Somewhat important; 2 - Quite important; 3 - Extremely important)

Imbalance in Treatment Group Size
1

Covariate Imbalance Z_1 (N(0,1))
1

Covariate Imbalance Z_2 (linear drift ranging from (-2,2] + N(0,1))
1

Covariate Imbalance Z_3 (autocorrelated)
1

Total Response Y_An_A + Y_Bn_B
1

Expected Number of Failures
1

Selection Bias
1

Accidental Bias
1

Relative Bias (naive)
1

Relative Bias (adjusted)
1

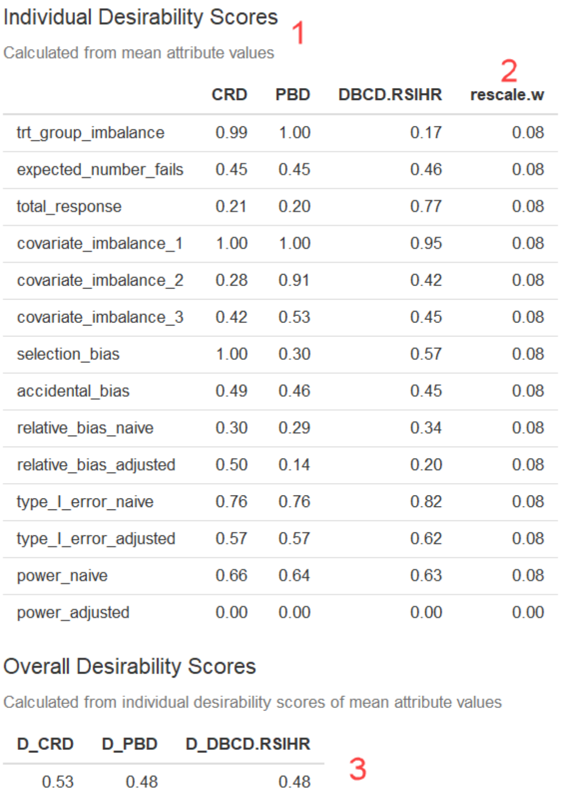
Type I Error (naive)
1

Type I Error (adjusted)
1

Power (naive)
1

Power (adjusted)
0

Figure B.19



Individual Desirability Scores

Calculated from mean attribute values

	CRD	PBD	DBCD.RSIHR	rescale.w
trt_group_imbalance	0.99	1.00	0.17	0.08
expected_number_fails	0.45	0.45	0.46	0.08
total_response	0.21	0.20	0.77	0.08
covariate_imbalance_1	1.00	1.00	0.95	0.08
covariate_imbalance_2	0.28	0.91	0.42	0.08
covariate_imbalance_3	0.42	0.53	0.45	0.08
selection_bias	1.00	0.30	0.57	0.08
accidental_bias	0.49	0.46	0.45	0.08
relative_bias_naive	0.30	0.29	0.34	0.08
relative_bias_adjusted	0.50	0.14	0.20	0.08
type_I_error_naive	0.76	0.76	0.82	0.08
type_I_error_adjusted	0.57	0.57	0.62	0.08
power_naive	0.66	0.64	0.63	0.08
power_adjusted	0.00	0.00	0.00	0.00

Overall Desirability Scores

Calculated from individual desirability scores of mean attribute values

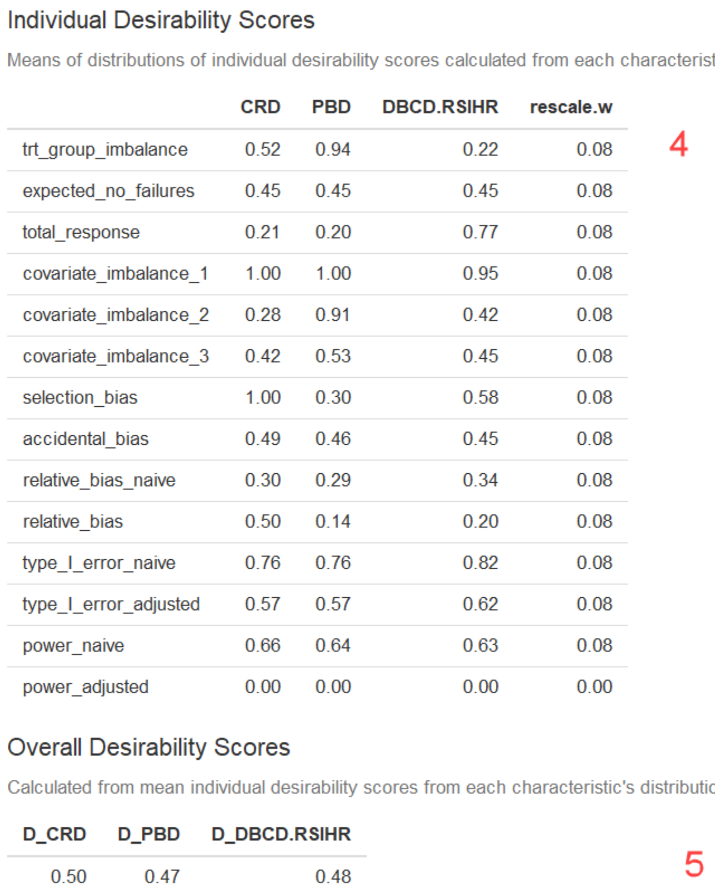
D_CRD	D_PBD	D_DBCD.RSIHR
0.53	0.48	0.48

Figure B.20

1 This is reiterating what was shown at the end of Step 2. It is a table of individual desirability scores for each characteristic, by design. These scores were calculated using a single mean value of the characteristic for each design.

2 This column **rescale.w** is the standardized weight as decided by your stated level of importance given to each characteristic. The weights are standardized so that they may add to 1. This will update in real-time as you change your preferences in the slider inputs above.

3 This table displays the Overall Desirability Scores of the designs using **1** and Equation B.11.



Individual Desirability Scores

Means of distributions of individual desirability scores calculated from each characteristic's distribution of realized values across specified iterations

	CRD	PBD	DBCD.RSIHR	rescale.w
trt_group_imbalance	0.52	0.94	0.22	0.08
expected_no_failures	0.45	0.45	0.45	0.08
total_response	0.21	0.20	0.77	0.08
covariate_imbalance_1	1.00	1.00	0.95	0.08
covariate_imbalance_2	0.28	0.91	0.42	0.08
covariate_imbalance_3	0.42	0.53	0.45	0.08
selection_bias	1.00	0.30	0.58	0.08
accidental_bias	0.49	0.46	0.45	0.08
relative_bias_naive	0.30	0.29	0.34	0.08
relative_bias	0.50	0.14	0.20	0.08
type_I_error_naive	0.76	0.76	0.82	0.08
type_I_error_adjusted	0.57	0.57	0.62	0.08
power_naive	0.66	0.64	0.63	0.08
power_adjusted	0.00	0.00	0.00	0.00

Overall Desirability Scores

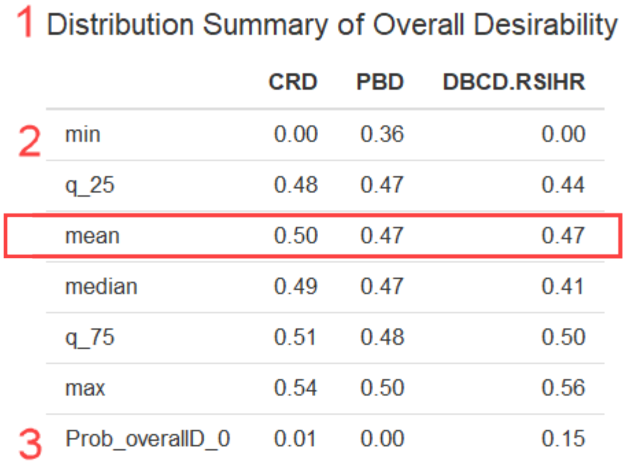
Calculated from mean individual desirability scores from each characteristic's distribution of individual desirability scores

D_CRD	D_PBD	D_DBCD.RSIHR
0.50	0.47	0.48

Figure B.21

4 This is also a reiteration of what was shown at the end of Step 2. This table summarizes the *mean* individual desirability score as calculated from the distribution of values of each characteristic which resulted from a number of simulated trials equaling the number of iterations requested by the user.

5 This table displays the Overall Desirability Scores of the designs using 4 and Equation B.11.



Distribution Summary of Overall Desirability

	CRD	PBD	DBCD.RSIHR
min	0.00	0.36	0.00
q_25	0.48	0.47	0.44
mean	0.50	0.47	0.47
median	0.49	0.47	0.41
q_75	0.51	0.48	0.50
max	0.54	0.50	0.56
Prob_overallD_0	0.01	0.00	0.15

Figure B.22

1 This table summarizes the distribution of overall desirability scores for each design.

2 The rows labeled min, q_25, mean, median, q_75, and max represent the minimum, 25th percentile, mean, median, 75th percentile, and maximum overall desirability scores for each design.

3 "Prob_overallD_0": Probability that the overall desirability score of the design will be 0, as estimated by the proportion of simulated trials where the design yielded an overall desirability score of 0.

Summary of Designs

Let $\mathcal{F}_n = T_1, ..., T_n$ be a set of treatment assignments for n stages of the randomization process.

We denote Treatment E as the experimental treatment, and Treatment C as the control treatment. A set of treatment assignments for n patients is $T_1, ..., T_n$, where $T_j = 1$ when patient j is assigned to treatment arm E, and $T_j = 0$ when patient j is assigned to treatment arm C. The probability of being assigned to Treatment E is denoted by $Pr(T_j = 1) = E(T_j)$.

Let Diff_n denote the difference in sample size between treatment group E and treatment group C. Specifically, let Diff_n = $n_E(n) - n_C(n)$.

Complete Randomization Design (CRD)

In Complete Randomization Design (CRD), each patient is enrolled into either treatment arm E or treatment arm C with probability 1/2. The allocation rule is independent of prior assignments:

$$E(T_j|\mathcal{F}_{j-1}) = E(T_j) = \frac{1}{2}. \tag{B.12}$$

There are no restrictions imposed upon this design.

Forced Balance Designs

Truncated Binomial Design (TBD)

The Truncated Binomial Design (TBD) (Blackwell and Hodges, 1957) is a forced balance procedure, meaning that exactly half of n patients will be assigned to each treatment arm. In this allocation rule, complete randomization is performed until one treatment arm contains half of the pre-determined sample size; subsequently, all remaining patients will receive the other treatment.

The truncated binomial design allocation rule is defined by:

$$\begin{aligned} E(T_j|\mathcal{F}_{j-1}) &= \frac{1}{2}, \text{if } \max(n_E(j-1), n_C(j-2)) < \frac{n}{2} \\ &= 0, \text{if } n_E(j-1) = \frac{n}{2} \\ &= 1, \text{if } n_C(j-1) = \frac{n}{2}. \end{aligned} \tag{B.13}$$

Random Allocation Rule (RAR)

The Random Allocation Rule (RAR) is also a forced balance procedure, with

$$E(T_j|\mathcal{F}_{j-1}) = \frac{\frac{n}{2} - n_E(j-1)}{n-(j-1)}, \qquad j = 2, ..., n, \tag{B.14}$$

and $E(T_1) = 1/2$.
One can think of this allocation rule in terms of an urn model. One samples from an urn with n/2 balls for treatment group E, and n/2 balls for treatment group C, without replacement.

Permuted Block Design (PBD)

In order to avoid severe treatment size imbalance during the entire course of a trial, clinical trialists often use “blocks”. Forced balance randomization within blocks is used in order to ensure balance at the end of

each block. Specifically, in the Permuted Block Design (PBD) (Zelen, 1974), there are M blocks of size B, where $B = n/M$. Each block is filled using a forced balanced procedure (e.g. Random Allocation Rule, Truncated Binomial Design), so that there are M occurrences of balanced allocation during the course of the trial. The maximum imbalance at any time point is then half a block size, $B/2$.
Let R_j define the position patient j takes within his block. If we fill blocks using RAR, the allocation rule is:

$$E(T_j|\mathcal{F}_{j-1}, B, R_j) = \frac{\frac{B}{2} - \sum_{l=j+1-R_j}^{j-1} T_l}{B - R_j + 1}. \tag{B.15}$$

Random Block Design (RBD.RAR, RBD.TBD)

The Random Block Design is a variation of the Permuted Block Design, with the difference lying in its ability to have multiple block sizes throughout the course of the trial. Block sizes are randomly selected from a discrete uniform distribution. Let B_max be the maximum treatment size difference, which is half of the largest block. The different block sizes, picked at random with probability $1/B_max$ after the fulfillment of a single block, are then 2, 4, 6, ..., $2B_max$. Let B_j be the block size of the block with the j^{th} patient. Let R_j define the position patient j takes within his block, ranging from 1,...,B_j. Each block can be filled with any forced balance procedure. If we fill each block using RAR, the allocation rule is:

$$E(T_j|\mathcal{F}_{j-1}, B_j, R_j) = \frac{\frac{B_j}{2} - \sum_{l=j+1-R_j}^{j-1} T_l}{B_j - R_j + 1}. \tag{B.16}$$

This website abbreviates Random Block Design using blocks filled with Truncated Binomial Design as RBD.TBD, and with Random Allocation Rule as RBD.RAR.

Biased Coin Designs

Efron's Biased Coin Design (BCD_p)

The Biased Coin Design (Efron, 1971) seeks to balance treatment assignments by allocating patients to the underrepresented treatment group with a higher probability. The allocation rule is defined as

$$E(T_j|\mathcal{F}_{j-1}) = \begin{cases} \frac{1}{2}, & \text{if } |\text{Diff}_{j-1}| = 0, \\ p, & \text{if } \text{Diff}_{j-1} < 0, \\ 1-p, & \text{if } \text{Diff}_{j-1} > 0, \end{cases} \tag{B.17}$$

where $0.5 < p <= 1$. It can be seen that when p = 1/2, we have complete randomization with the restriction of a maximal imbalance of n/2. At p = 1, Efron's biased coin design simplifies to a permuted block design with a block size of 2, so that every other patient's allocation assignment is deterministic, and maximal imbalance for even n being 0. The parameter p thus represents a trade-off between balance and predictability. Efron's original paper states:
"The value p = 2/3, which is the author's personal favourite, will be seen to yield generally good designs..."

Big Stick Design (BSD)

The Big Stick Design (BSD) allows a degree of imbalance up to a magnitude given by a fixed imbalance tolerance parameter b. The allocation rule is given by:

$$E(T_j|\mathcal{F}_{j-1}) = \begin{cases} \frac{1}{2}, & \text{if } |\text{Diff}_{j-1}| < b, \\ 0, & \text{if Diff}_{j-1} = b, \\ 1, & \text{if Diff}_{j-1} = -b \end{cases} \tag{B.18}$$

Big Stick Design(proportion) (proportionBSD)

The Big Stick Design with Maximum Proportionate Degree of Imbalance replaces the absolute difference used in the Big Stick Design with an acceptable degree of imbalance, $D_{j-1}/(j-1)$. The allocation rule, then, is:

$$E(T_j|\mathcal{F}_{j-1}) = \begin{cases} \frac{1}{2}, & \text{if Diff}_{j-1}/(j-1) < \text{prop}, \\ 0, & \text{if Diff}_{j-1}/(j-1) = \text{prop}, \\ 1, & \text{if Diff}_{j-1}/(j-1) = -\text{prop}, \end{cases} \tag{B.19}$$

where prop is a pre-defined acceptable degree of imbalance.

Biased Coin Design with Imbalance Intolerance (BCD2_p)

The Biased Coin Design with Imbalance Intolerance (BCDII(p)) combines the concepts of the big stick design and Efron's biased coin design (Chen, 1999). The allocation rule is defined by:

$$E(T_j|\mathcal{F}_{j-1}) = \begin{cases} \frac{1}{2}, & \text{if Diff}_{j-1} = 0, \\ 0, & \text{if Diff}_{j-1} = b, \\ 1, & \text{if Diff}_{j-1} = -b, \\ p, & \text{if } 0 < \text{Diff}_{j-1} < b, \\ 1-p, & \text{if } -c < \text{Diff}_{j-1} < 0 \end{cases} . \tag{B.20}$$

Note that b is the maximum tolerable imbalance, and p is the probability of assigning patients to the experimental arm E.

Accelerated Biased Coin Design (ABCD_a)

The Accelerated Biased Coin Design (ABCD(a)) is a bigger umbrella of which the big stick design, Efron's biased coin design, and biased coin design with imbalance intolerance are special cases. Let F be a function that maps integers to [0,1] such that $F(x)$ is decreasing, and $F(-x) = 1 - F(x)$. The allocation rule is then:

$$E(T_j|\mathcal{F}_{j-1}) = F(\text{Diff}_{j-1}),$$

where

$$F_a(x) = \begin{cases} \frac{|x|^a}{|x|^a+1}, & \text{if } x \leq -1, \\ \frac{1}{2}, & \text{if } x = 0, \\ \frac{1}{|x|^a+1}, & \text{if } x \geq 1. \end{cases} \tag{B.21}$$

The a parameter controls the degree of randomness, with $a = 0$ equating complete randomization. As $a \to \infty$, the ABCD is equivalent to the Big Stick Design with $b = 2$.

Response-Adaptive Randomization (RAR)

Response-Adaptive Randomization (RAR) designs adapt to the responses observed in the study, increasing or decreasing the probability of a subject being assigned to a given treatment arm. RAR designs target an allocation proportion, which depends on pre-stated objectives. Section B provides an overview of target allocations frequently discussed in literature.

Target Allocations

Target allocation is defined as the ideal proportion $n_E/(n_E + n_C)$. The ideal proportion varies depending on one's objectives. The target allocations available in this website are summarized in Table B.2.

Objective	Allocation Name	Binary	Continuous Normal
1. Maximize power for a fixed sample size	Neyman	$\frac{\sqrt{p_E q_E}}{\sqrt{p_E q_E}+\sqrt{p_C q_C}}$	$\frac{\sigma_E}{\sigma_E+\sigma_C}$
2. Minimize expected number of treatment failures for a fixed power	RSIHR	$\frac{\sqrt{p_E}}{\sqrt{p_E}+\sqrt{p_C}}$	$\frac{\sqrt{\mu_C}\sigma_E}{\sqrt{\mu_C}\sigma_E+\sqrt{\mu_E}\sigma_C}$
3. Minimize treatment failures and ensure fewer patients are allocated to inferior treatment	RSIHR2	NA	$\frac{\sqrt{\mu_C}\sigma_E}{\sqrt{\mu_C}\sigma_E+\sqrt{\mu_E}\sigma_C}$ if ($\mu_E < \mu_C$ and $\sigma_E\sqrt{\mu_C}/\sigma_C\sqrt{\mu_E}) > 1$ or ($\mu_E > \mu_C$ and $\sigma_E\sqrt{\mu_C}/\sigma_C\sqrt{\mu_E}) < 1$; 1/2 otherwise
4. Urn model (good for lowering expected number of treatment failures when $p_E + p_C > 1$)	Urn	$\frac{q_C}{q_E+q_C}$	NA
5. Minimize patients with response greater than c	Biswas Mandal (BM)	NA	$\frac{\sqrt{\Phi(\frac{\mu_C-c}{\sigma_C})}\sigma_E}{\sqrt{\Phi(\frac{\mu_C-c}{\sigma_C})}\sigma_E+\sqrt{\Phi(\frac{\mu_E-c}{\sigma_E})}\sigma_C}$
6. Not formally defined	Bandyopadhyay Biswas (BB)	NA	$\Phi(\frac{\mu_C-\mu_E}{T})$
7. Minimize the maximum eigenvalue of the inverse of Fisher's information (E-optimality)	Baldi Antognini Giovagnoli (Baldi)	see text	NA

Table B.2: Summary of Allocations Targeted By RAR Designs.

Neyman Allocation (Neyman)

The Neyman allocation maximizes power for a given sample size N and fixed probabilities of success. The Neyman allocation can be applied to either binary or continuous responses.

$$\varphi = \frac{\sqrt{p_E q_E}}{\sqrt{p_E q_E} + \sqrt{p_C q_C}}.$$

This is known as the Neyman allocation. A weakness of this allocation is that when the success probabilities on both treatment groups are high ($p_E + p_C > 1$), more subjects are assigned to the weaker treatment arm.

RSIHR Allocation (RSIHR)

While one objective could be to minimize the total sample size in a trial while still achieving sufficient power, a second objective could be instead to minimize the total number of treatment failures. RSIHR allocation (named after the initials of the authors on the original paper) seeks to minimize treatment failures for a fixed power, and is given by:

$$\varphi = \frac{\sqrt{p_E}}{\sqrt{p_E} + \sqrt{p_C}}.$$

In the continuous case,

$$\varphi = \frac{\sqrt{\mu_C}\sigma_E}{\sqrt{\mu_C}\sigma_E + \sqrt{\mu_E}\sigma_C}.$$

However, when $\mu_E < \mu_C$, it is possible for $\frac{n_E}{n_C}$ to be less than 1/2. This shows that while power is maximized for a fixed expected number of treatment failures, this target allocation has the potential to allocate more patients to the inferior treatment. RSIHR2 seeks to remove this ethical flaw by modifying the allocation.

Urn Allocation (Urn)

In trials with binary responses, urn allocation can be used when the probability of success is high in both the experimental and the control arms, specifically, when $p_E + p_C > 1$. The target allocation is

$$\varphi = \frac{q_E}{q_E + q_C}.$$

Urn models such as the randomized play-the-winner (RPW) (Wei et al., 1978) and drop-the-loser (DL) rule (Ivanova, 2003) are procedures that target the Urn allocation, which is only a property of the procedure rather than the solution to a specific optimality problem.

Biswas and Mandal Allocation (BM)

Biswas and Mandal (2004) sought to generalize binary optimal allocation in terms of failures to one for normal responses. This allocation assumes that smaller responses are better and minimizes the total number of patients with response greater than c, thereby minimizing the number of failures as given by a threshold. The target allocation is given by

$$\varphi = \frac{\sqrt{\Phi(\frac{\mu_C - c}{\sigma_C})}\sigma_E}{\sqrt{\Phi(\frac{\mu_C - c}{\sigma_C})}\sigma_E + \sqrt{\Phi(\frac{\mu_E - c}{\sigma_E})}\sigma_C}.$$

The website ascertains your preference for the value of c with the question: “This question pertains only if smaller values of the outcome are considered better. What is the maximum acceptable value of the outcome? Values above this will be counted as failures.”

Bandyopadhyay and Biswas Allocation (BB)

Bandyopadhyay and Biswas (2001) proposed a target allocation that does not seek to optimize any formal objective property. T_{BB} is a scaling factor and is set at 2 in this website as is suggested in the original paper. The target allocation is given by:

$$\varphi = \Phi(\frac{\mu_C - \mu_E}{T_{BB}}).$$

Baldi Antognini and Giovagnoli (Baldi)

Baldi Antognini and Giovagnoli (2010) proposed a target allocation that has both ethical and inferential aims. They seek to minimize the maximum eigenvalue of the inverse of Fisher's information (E-optimality). Consider the compound criterion which combines ethical and inferential objectives:

$$\Phi_w(\varphi) = w\left(\frac{\psi_E(\varphi)}{\psi_E^*}\right) + (1-w)\left(\frac{\psi_I(\varphi)}{\psi_I^*}\right),$$

where $w \in (0,1)$ is a user-defined weight for importance of ethics, $1-w$ is the weight given to inference, and $\psi_E(\varphi) = q_E\varphi + q_C(1-\varphi)$ is the expected proportion of treatment failures, $\psi_E^* = min(q_E, q_C)$, $\psi_I(\varphi) = p_E q_E/\varphi + p_C q_C/(1-\varphi)$ is the variance of the estimated treatment difference, and $\psi_I^* = (\sqrt{p_E q_E} + \sqrt{p_C q_C})^2$ is the minimum value of $\psi_I(\varphi)$ for $\varphi \in (0,1)$. The goal is to minimize the compound criterion. We can see, then, w places more importance on minimizing $\psi_E(\varphi)$, the expected proportion of failures, and $(1-w)$ places more importance on $\psi_I(\varphi)$, which minimizes the variance of the estimated treatment difference.
The target allocation φ is the solution in (0,1) of the following equation

$$\frac{w}{1-w}\frac{p_E - p_C}{\min(q_E, q_C)}\left(\frac{\sqrt{p_C q_C}}{\sqrt{p_E q_E}} + 1\right)^2 = \frac{(\frac{\sqrt{p_C q_C}}{\sqrt{p_E q_E}} - 1)\varphi^2 + 2\varphi - 1}{(\varphi(1-\varphi))^2}.$$

Response-Adaptive Randomization (RAR) Designs

Eisele & Woodroofe, 1995(EW1995)

Eisele & Woodroofe (1995) presented a response-adaptive randomization (RAR) procedure. If $T_j = 1$, X_j is normally distributed with mean μ_E and variance σ_E^2. If $T_j = 0$, X_j is normally distributed with mean μ_C and variance σ_C^2.
Let $g(x,y) = [1 - (\frac{1}{y} - 1)x]$.
Then Eisele & Woodroofe's procedure is defined by:

$$\phi = g\left(\frac{n_E(j-1)}{j-1}, \varphi(\hat{\theta}_j)_{\text{target}}\right). \tag{B.22}$$

Doubly-Biased Coin Design (DBCD)

The Doubly Biased Coin Design (DBCD) is a RAR procedure that obtained its name due to its consideration of both the proportion of enrolled patients assigned to each treatment arm and the estimate of the target allocation proportion. Biased coin designs are able to reduce experimenter/selection bias. The procedure aims to fulfill the goal of allocating n_E patients to treatment E, such that $\frac{n_E}{n_E+n_C}$ equals the target allocation proportion. Treatment E is assigned, then, with a probability less than the current maximum likelihood estimate (MLE) of the target proportion when the observed proportion is larger than this estimate. Similarly, Treatment E is assigned with a probability greater than the current MLE of the target proportion when the observed proportion is larger than this estimate.
In the DBCD, the first $2m_0$ patients are enrolled with the probability of being assigned to treatment E equal to $probn_E = \frac{m_0 - n_E}{2m_0 - (j-1)}$, where j the patient number. After $2m_0$ patients have been enrolled, the probability of being assigned to treatment arm E is determined by

$$\phi = \frac{\frac{R}{R+1}(\frac{\frac{R}{R+1}}{\frac{n_E}{N}})^\gamma}{\frac{R}{R+1}(\frac{\frac{R}{R+1}}{\frac{n_E}{N}})^\gamma + (1 - \frac{R}{R+1})(\frac{1-\frac{R}{R+1}}{1-\frac{n_E}{N}})^\gamma}, \tag{B.23}$$

where R is recalculated for each enrolling patient using estimates of the means and variances of the previously enrolled patients, and where $\gamma >= 0$ determines the degree of randomness, with $\gamma = 0$ being the most random and γ −¿ ∞ being an almost deterministic procedure. Correlation is held constant and assumed known throughout the study.

Sequential Maximum Likelihood Estimation Design (SMLE)

The Sequential Maximum Likelihood Estimation (SMLE) Design sets treatment randomization probabilities to be equal to the current estimates of the target allocation proportions. It is equivalent to DBCD with $\gamma = 0$. This design can lead to a modest reduction in treatment failures with minimal loss in power relative to equal randomization designs. However, SMLE has also been shown to be quite variable, with potential negative effects on power.

Efficient Randomized-Adaptive Design (ERADE)

Efficient Randomized-Adaptive Design (ERADE) (Hu, et al. YEAR) is an extension of the biased-coin design; it is equivalent to DBCD with $\gamma \to \infty$. ERADE is a RAR procedure that can target any pre-specified allocation proportion, while still preserving allocation randomness and boasting minimal variability. The theoretical properties of ERADE echo those of DBCD: both resulting sample proportions (n_E/n) and estimators are strongly consistent and asymptotically normal.

$$\phi = \begin{cases} \delta\hat{\rho}_(j-1) & \text{if } n_E(j-1)/n > \hat{\rho}_(j-1) \\ \hat{\rho}_(j-1) & \text{if } n_E(j-1)/n = \hat{\rho}_(j-1) \\ 1-\delta(1-\hat{\rho}_(j-1)) & \text{if } n_E(j-1)/n < \hat{\rho}_(j-1) \end{cases} \tag{B.24}$$

The discrete property of ERADE differentiates it from DBCD, resulting in less variability.

Evaluated Characteristics of Designs

1. Treatment Group Size Imbalance
2. Expected Number of Failures
3. Expected Total Response (continuous responses only)
4. Accidental Bias
5. Covariate Imbalance (Z_1)
6. Covariate Imbalance (Z_2)
7. Covariate Imbalance (Z_3)
8. Selection Bias
9. Relative Bias (Naive)
10. Relative Bias (Adjusted)
11. Type I error in the Presence of Chronological Bias (Naive)
12. Type I error in the Presence of Chronological Bias (Adjusted)
13. Power in the Presence of Chronological Bias (Naive)
14. Power in the Presence of Chronological Bias (Adjusted)

Treatment Group Size Imbalance

Definition: $n_E - n_C$.

Complete randomization and restricted randomization procedures are used for balancing treatment assignments. Equal allocation amongst treatment groups has two benefits: first, equal allocation caters to the ethical concept of equipoise that clinical trialists should believe to be true at the start of a trial. Second, balanced group sizes in clinical trials hedge against accidental bias, the bias in the estimated treatment effect that is due to the omission of significant confounders from the model. In traditional (non-adaptive) designs, power typically is maximized when the samples sizes in the treatment arms are the same.
For each design evaluated, then, treatment arm imbalance is assessed at the end of each simulated iteration. The distribution of treatment arm imbalance for each given design is then assessed.

Expected Number of Failures
Definition: $f_E + f_C$.
In recent adaptive designs, the ethical objective of minimizing the number of failures in a trial has been a characteristic under scrutiny. When the outcome is binary, the outcome is already "success" or "failure", making this calculation straight forward. When the outcome is continuous and considered smaller the better, a failure is defined as any outcome greater than a pre-defined maximum threshold. The program tallies the number of failures in each simulated clinical trial and stores the distribution. Let iter = the total number of iterations (the total number of simulated trials), and f_{E_i}, f_{C_i} be the number of failures in the experimental and control arms in iteration i, respectively. Then the number of failures across iterations is stored:

$$\begin{pmatrix} f_{E_1} & f_{C_1} \\ f_{E_2} & f_{C_2} \\ \vdots & \vdots \\ f_{E_i} & f_{C_i} \\ \vdots & \vdots \\ f_{E_{\text{iter}}} & f_{C_{\text{iter}}} \end{pmatrix}.$$

The Average Expected Number of Failures of a design is calculated as the average number of failures in the design over a large number of iterations. Then the

$$\text{Average Expected Number of Failures } = \frac{\sum_{i=1}^{\text{iter}} f_{E_i} + f_{C_i}}{\text{iter}}. \tag{B.25}$$

You will be able to evaluate a design by considering both the entire distribution of expected number of failures, as well as by just the average expected number of failures. If this characteristic of a design is important to you, consider using a response-adaptive-randomization design (ERADE, DBCD, SMLE, EW1995) that targets the RSIHR allocation.

Expected Total Response
Definition: $\bar{Y}_E n_E + \bar{Y}_C n_C$.
This characteristic is evaluated for continuous responses only. You will be asked if smaller responses are better, larger responses are better, or responses within a target value are better. If smaller responses are better, and minimizing the total response is important to you, consider using a response-adaptive randomization design (ERADE, DBCD, SMLE, EW1995) that targets the RSIHR allocation. For smaller-the-better responses, you will be asked if you suspect correlation is present between responses of the experimental and control arms due to some common exposure (e.g. all subjects treated in same hospital). If correlation is suspected, an additional allocation - R.corr - will also be targeted.

Accidental Bias
Definition:

$$\text{Accidental Bias Factor Estimate } = \left(\frac{n}{n^2 - (e't)^2}\right)^2 \lambda_{\hat{max}}, \tag{B.26}$$

where λ_{max} is the maximum value of the covariance matrix of T. Note that in this overview of accidental bias, we shall modify notation and allow T_j = 1 for an individual in treatment arm E, and T_j = -1 for an individual in treatment arm C, $j = 1, ..., n$. The covariance matrix of T is:

$$VarT = \Sigma = \begin{bmatrix} E[(T_1 - ET_1)(T_1 - ET_1)] & E[(T_1 - ET_1)(T_2 - ET_2)] & \dots & E[(T_1 - ET_1)(T_n - ET_n)] \\ E[(T_2 - ET_2)(T_1 - ET_1)] & E[(T_2 - ET_2)(T_2 - ET_2)] & \dots & E[(T_2 - ET_2)(T_n - ET_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(T_n - ET_n)(T_1 - ET_1)] & E[(T_n - ET_n)(T_2 - ET_2)] & \dots & E[(T_n - ET_n)(T_n - ET_n)] \end{bmatrix},$$

where the expected value of a treatment for patient j is estimated by taking the mean of the patient's treatment indicator value ($T_j = -1$ or 1) across all iterations. Specifically, for patient j and iteration i ($i = 1, ..., \text{iter}$), $ET_j = \frac{\sum_{i=1}^{\text{iter}} T_j}{\text{iter}}$. This estimate of ET_j then allows us to find the estimate $\hat{\Sigma}$, which is then used to find an estimate of the accidental bias factor.

Accidental bias describes the measure of bias in the treatment effect that is introduced due to some unobserved yet confounding covariate. Specifically, if β is the true value of the coefficient of a confounder, then the bias of the treatment effect is equal to the accidental bias factor multiplied by β^2. See Appendix or Lachin & Rosenberger (2016) for more details.

As the number of iterations $iter \to \infty$, the estimate of the accidental bias factor reaches its theoretical value. Designs yielding lower accidental bias factor estimates are favorable. In practice, this means the designs with lower average treatment group imbalance, and with less variability within treatment assignments for each subject j, are less likely to have treatment effect estimates impacted by unobserved covariates.

Lachin et al. note that with the exception of truncated binomial design, accidental bias seems to be negligible for forced balanced designs. However, accidental bias in response-adaptive designs seems to be an area less studied. Using the accidental bias factor estimate from Equation B.26, this website is able to compare the impact of unobserved covariates on a broader range of designs.

Covariate Imbalance ($Z1, Z2, Z3$)
Definition:

$$\text{Probability}(\left|\overline{Z}_E - \overline{Z}_C\right| > \epsilon) = \frac{\sum_{i=1}^{\text{iter}} \mathbb{1}(\left|\overline{Z}_{E_i} - \overline{Z}_{C_i}\right| > \epsilon)}{\text{iter}}$$

The setup for the study of covariate imbalance follows that of Lachin and Rosenberger (2016). Three different patterns are studied to see the probability of imbalance at the end of a trial:
1. $Z1_1, Z1_2, ..., Z1_n$ are i.i.d. N(0,1).
2. $Z2_1, Z2_2, ..., Z2_n$ drift linearly over time on the interval (-2,2] + a N(0,1) random variable.
3. $Z3_1, Z3_2, ..., Z3_n$ are autocorrelated. Specifically, $Z3_j = Z3_{j-1}$ + a N(0,1) random variable, with j = 2,...,n and the first of the series $Z3_1$ equaling a N(0,1) random variable.

These three scenarios represent three different types of covariates. The first is standardized normal, and is a good representation of what one expects from most covariates measured in a study. Different means and variances can be simulated, but N(0,1) is a representative proxy. The second scenario is representative of a covariate subject to a linear time trend. This is not to be confused with a linear time trend influencing the primary outcome of interest Y of a trial. An example of a covariate subject to a linear drift may be improving average blood pressure in patient population, when blood pressure is not the primary endpoint of interest. The third scenario is for autocorrelated variables, also known as *serially correlated* or *serially dependent*. This means that a covariate value is not independently and identically distributed, but rather, depends on the previous value. Returns on stock prices are frequently used as an example of autocorrelated variables, since past returns seem to influence future performance and returns. Other examples include annual rainfall, sunspot activity, and the price of agricultural products. In health, autocorrelation is seen in covariates quantifying exposure to pollutants. For example, asthma symptoms and daily ambient particulate matter concentrations are characterized as being related through an autocorrelated lag model. At the end of each trial, we compute the frequency in which $|\overline{Z}_E - \overline{Z1}_C| > \epsilon$, where ϵ can be specified by the user. The (frequency/total number of iterations) yields the simulated probability of covariate imbalance for these three scenarios.

Selection Bias
Definition:

$$\rho_{pred} = \sum_{j=1}^{n} E \left| [E(T_j|F_(j-1))] - \frac{1}{2} \right| .$$

ρ_{pred} is the predictability of a randomization sequence, as measured by the difference between the conditional probability of treatment assignment and the conditional probability. This is a measure of third-order selection bias, which occurs when only future patient allocations are concealed, thus giving the investigator the ability to predict future allocations based on prior assignments. While in today's clinical trials, a clinical trial usually targets double-blinded treatment assignment (the individual responsible for assigning treatments and the analysts do not know who receives which treatment), in reality selection bias is still a risk that should be considered during the design of a trial. In spite of the double-blinded nature of many trials, treatment assignment may sometimes still be guessed or even obvious: patients in different treatment arms may experience different side effects; or sometimes the treatment arms themselves are unable to be masked (e.g. surgery vs. chemo). In a meta-analysis of randomized clinical trials in which surgery was an intervention, less than 25 % of the trials concealed treatment allocation. The literature has shown that systematic baseline imbalances can occur in randomized trials, where selection bias would force imbalance in covariates influencing patient allocation. Berger (2005) states that randomization is necessary to ensure that any observed baseline imbalances are random, but it is not sufficient.

If the intervention of your trial is hard to mask (e.g. surgical, behavioural), consider selection bias during the design of your study. If you are certain that third-order selection bias is not relevant to your trial, you may ignore this characteristic by telling the program Selection Bias is "Not important at all" by giving it a score of 0 in Step 3.

Relative Bias
Relative Bias is a useful measure to see the bias of the treatment effect in the simulated trials. One can see that certain designs yield less biased treatment effect effects. The relative bias is defined as $\frac{\hat{\beta_1}-\beta}{\beta} \times 100$. The naive relative bias is based off of $\hat{\beta_1}$ from Equation B.6, while the adjusted estimate is based off of $\hat{\beta_1}$ from Equation B.7.

Type I error in the Presence of Chronological Bias
This is the proportion of times the simulated trials result in a rejection of the null hypothesis when in fact the null hypothesis is true. The naive Type I error results from an analysis from Equation B.6, whlie the adjusted Type I error results from analysis from Equation B.7.

Power in the Presence of Chronological Bias Power is the proportion of times the simulated trials result in the correct rejection of the null hypothesis when indeed the alternative hypothesis is true. The naive power results from an analysis from Equation B.6, whlie the adjusted power results from analysis from Equation B.7.

Accidental Bias Factor
Let us consider the true model to be a standard normal error regression model:

$$E(Y) = \mu e + \alpha t + \beta z,$$

where e is a vector of ones: $e = (1, 1, ..., 1)'$, t is the treatment vector given by $t = T = (t_1, ..., t_n)'$, and z is a covariate that is significantly associated with the outcome Y. Note that in this setup, α is the coefficient for the treatment effect.
Denoting the design matrix X, we see that

$$X = \begin{bmatrix} 1 & t_1 & z_1 \\ 1 & t_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & z_n \end{bmatrix}, \qquad X'Y = \begin{bmatrix} e'Y \\ t'Y \\ z'Y \end{bmatrix}.$$

Using ordinary least squares method, if we look at $(X'X)^{-1}X'Y$, then the consistent estimate of α is

$$E(\hat{\alpha}) = \frac{n(\mu e't + n\alpha + \beta z't) - e't(n\mu + \alpha e't)}{n^{-}(e't)^2}.$$

However, if the covariate z is incorrectly excluded from the model,

$$E(Y) = \mu e + \alpha t,$$

Denoting the design matrix X, we see that

$$X = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}, \qquad X'Y = \begin{bmatrix} e'Y \\ t'Y \end{bmatrix}$$

.

$$(X'X)^{-1} = \frac{1}{n^2 - (e't)^2} \begin{bmatrix} n & -e't \\ -e't & n \end{bmatrix}$$

.
Using ordinary least squares method, if we look at $(X'X)^{-1}X'Y$, then the biased estimate is

$$\hat{\alpha} = \frac{nt'Y - (e't)(e'Y)}{n^2 - (e't)^2}.$$

The squared bias term is then

$$[E(\hat{\alpha} - \alpha)]^2 = (\frac{n}{n^2 - (e't)^2})^2 \beta^2 (z't)^2$$

.
The impact of imbalanced treatment group sizes is highlighted by the $(e't)$ and is clear: larger imbalances contribute to greater bias in the estimate of the treatment effect. The bias in the estimate of the treatment effect also increases with the magnitude of the coefficient β for the omitted covariate z. Lastly, accidental bias depends on the term $(z't)^2$, which is zero when z is orthogonal to t. The unconditional expectation can be taken for a fixed vector z, with t being a realization of T and $\Sigma_T = Var(T)$:

$$E(z'T)^2 = z'\Sigma_T z,$$

By Rao, $E(z'T)^2$ cannot exceed the maximum eigenvalue of Σ_T (Rao 1973 p62). Due to this inequality, Efron uses the maximum eigenvalue of Σ_T as a criterion to evaluate the degree to which accidental bias impacts a design.

Bibliography

[1] Aggarwal, A., Singh, H., Kumar, P., and Singh, M. Optimization of multiple quality characteristics for cnc turning under cryogenic cutting environment using desirability function. *Journal of Materials Processing Technology 205*, 1 (2008), 42–50.

[2] Alexander, F., Anderson, T., Brown, H., Forrest, A., Hepburn, W., and A.E., K. 14 years of follow-up from the edinburgh randomised trial of breast-cancer screening. *Lancet 353* (1999), 1903–1908.

[3] Altman, D. G., and Royston, J. P. The hidden effect of time. *Statistics in Medicine 7* (1988), 629–637.

[4] Antognini, A., and Giovagnoli, A. A new 'biased coin design' for the sequential allocation of two treatments. *Journal of the Royal Statistical Society 53* (2004), 651–664.

[5] Antognini, A., and Giovagnoli, A. Compound optimal allocation for individual and collective ethics in binary clinical trials. *Biometrika 97* (2010), 935–946.

[6] Atkinson, A. C. Selecting a biased-coin design. *Statistical Science 29*, 1 (2014), 144–163.

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

[8] Berger, V., Ivanova, A., and Knoll, M. Minimizing predictability while retaining balance through the use of less restrictive randomization procedures. *Statistics in Medicine 22*, 19 (2003), 3017–28.

[9] Berger, V. W. *Selection Bias and Covariate Imbalances in Randomized Clinical Trials.* John Wiley & Sons, 2005.

[10] Berger, V. W. Minimization, by its nature, precludes allocation concealment, and invites selection bias. *Contemporary Clinical Trials 31*, 5 (2010), 406.

[11] Biswas, A., and Bhattacharya, R. Optimal response-adaptive allocation designs in phase iii clinical trials: Incorporating ethics in optimality. *Statistics and Probability Letters 81* (2011), 1155–1160.

[12] Biswas, A., and Coad, D. S. A general multi-treatment adaptive design for multivariate responses. *Sequential Analysis Design Methods and Applications 24*, 2 (2005), 139–158.

[13] Biswas, A., and Mandal, S. Optimal adaptive designs in phase iii clinical trials for continuous responses with covariates. *m0Da 7 - Advances in Model-Oriented Design and Analysis 7* (2004), 51–58.

[14] Biswas, A., and Mandal, S. Optimal allocation proportion for a two-treatment clinical trial having correlated binomial responses. *mODa9 - Advances in Model-Oriented Design and Analysis 9* (2010), 41–48.

[15] Blackwell, D., and Hodges, J. Design for the control of selection bias. *Annals of Mathematical Statistics 28*, 2 (1957), 449–460.

[16] Bourguignon, B., and Massart, D. L. Simultaneous optimization of several chromatographic performance goals using derringer's desirability function. *Journal of Chromatography 586*, 1 (1991), 11–20.

[17] Chen, H., Wong, W., and Xu, H. An augmented approach to the desirability function. *Journal of Applied Statistics 39*, 3 (2012), 599–613.

[18] Chen, H., Wong, W., and Xu, H. Data-driven desirability function to measure patients' disease progression in a longitudinal study. *Journal of Applied Statistics* (2015).

[19] Chen, Y.-P. Biased coin design with imbalance tolerance. *Communications in Statistics. Stochastic Models 15*, 5 (1999).

[20] Ch'ng, C., Quah, S., and Low, H. A new approach for multiple-response optimization. *quality Engineering 17*, 4 (2005), 621–626.

[21] Coffey, T., Gennings, C., and Moser, V. The simultaneous analysis of discrete and continuous outcomes in a dose–response study: using desirability functions. *Regulatory Toxicology and Pharmacology 48*, 1 (2007), 51–58.

[22] Connor, E., Sperling, R., Gelber, R., Kiselev, P., Scott, G., OSullivan, M., VanDyke, R., Bey, M., Shearer, W., Jacobsen, R., and et al. Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. pediatric aids clinical trials group protocol 076 study group. *The New England Journal of Medicine 331* (1994), 1173–1180.

[23] Cruz-Monteagudo, M., Borges, F., and Cordeiro, M. N. D. Desirability-based multiobjective optimization for global qsar studies: Application to the design of novel nsaids with improved analgesic, antiinflammatory, and ulcerogenic profiles. *Journal of Computational Chemistry 29*, 14 (2008), 2445–2459.

[24] Dalkey, N., and Helmer, O. An experimental application of the delphi method to the use of experts. *Management Science 9*, 3 (1963), 351–515.

[25] Del Castillo, E., Montgomery, D., and McCarville, D. Modified desirability functions for multiple response optimization. *Journal of Quality Technology 28* (1996), 337–345.

[26] Derringer, G., and Suich, R. Simultaneous optimization of several response variables. *Journal of Quality Technology 12*, 4 (1980), 214–219.

[27] Dewe, W., Durand, C., Marion, S., Oostvogels, L., Devaster, J.-M., and Fourneau, M. A multi-criteria decision making approach to identify a vaccine formulation. *Journal of Biopharmaceutical Statistics 26*, 2 (2016), 352–364.

[28] Efron, B. Forcing a sequential experiment to be balanced. *Biometrika 58*, 3 (1971), 403–417.

[29] Eisele, J. The doubly adaptive biased coin design for sequential clinical trials. *Mournal of Statistical Planning and Inference 38* (1971), 249–261.

[30] Follmann, D., and Proschan, M. On the design and analysis of clinical trials with correlated outcomes. *Contemporary Clinical Trials 39*, 1 (2014), 86–94.

[31] Fransen, J., Kavanaugh, A., and Borm, G. Desirability scores for assessing multiple outcomes in systemic rheumatic diseases. *Communication in Statistics - Theory and Methods 38*, 18 (2009), 3461–3471.

[32] Friedman, L., Furberg, C., and DeMets, D. *Fundamentals of Clinical Trials.* Wright PSG, 1981.

[33] Gennings, C. Use of desirability functions to evaluate health status in patients with cirrhosis. *Journal of Hepatology 52*, 5 (2010), 665–671.

[34] Goldberg, D. E. *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley Publishing Company, 1989.

[35] Govaerts, B., and Tilleghem, C. Uncertainty propagation in multiresponse optimization using a desirability index. *No. STAT Discussion Paper (0532). UCL* (2005).

[36] Grantcharov, T., Kristiansen, V., Bendix, J., Bardram, L., Rosenberg, J., and Funch-Jensen, P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *BJS 91*, 2 (2004), 146–150.

[37] Group, I. E. W. Statistical principles for clinical trials e9.

[38] Group, T. A. W. Clopidogrel plus asprin versus oral anticoagulation for atrial fibrillation in the atrial fibrillation clopidogrel trial with irbesartan for prevention of vascular events (active w): a randomised controlled trial. *The Lancet 367*, 9526 (2006), 1903–1912.

[39] Hanin, L. Why statistical inference from clinical trials is likely to genreate false and irreproducible results. *BMC Medical Research Methodology 17*, 127 (2017).

[40] Harrington, E. C. The desirability function. *Industrial Quality Control 21*, 10 (1965), 494–498.

[41] He, W., Pinheiro, J., and Kuznetsova, O. M. *Practical Considerations for Adaptive Trial Design and Implementation.* Springer, 2014.

[42] He, Z., Zhu, P.-F., and Park, S.-H. A robust desirability function method for multi-response surface optimization considering model uncertainty. *European Journal of Operational Research 221*, 1 (2012), 241–247.

[43] Hemming, K., Taljaard, M., and Forbes, A. Modeling clustering and treatment effect heterogeneity in parallel and stepped-wedge cluster randomized trials. *Statistics in Medicine 37* (2017), 883–898.

[44] Holland, J. H. Genetic algorithms: Computer programs that evolve in ways that resemble natural selection can solve complex problems even their creators do not fully understand, 2005. `https://www.cc.gatech.edu/~turk/bio_sim/articles/genetic_algorithm.pdf`, accessed 21 November 2017.

[45] Honari, M., Askari, H., and Khosrowchahli, M. Use of desirability function method in optimization of regeneration and callus induction of alhagi camelorum. *American Journal of Plant Sciences 5* (2014), 268–274.

[46] Hu, F., and Rosenberger, W. F. *The Theory of Response-Adaptive Randomization in Clinical Trials.* John Wiley & Sons, 2006.

[47] Hu, F., and Zhang, L.-X. Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. *Annals of Statistics 32*, 1 (2004), 268–301.

[48] Hu, F., Zhang, L.-X., and He, X. Efficient randomized-adaptive designs. *The Annals of Statistics 37* (2009).

[49] Jacquier, I., Boutron, I., Moher, D., Roy, C., and Ravaud, P. The reporting of randomized clinical trials using a surgical intervention is in need of immediate improvement. *Annals of Surgery 244*, 5 (2006), 677–683.

[50] Jennison, C., and Turnbull, B. *Group Sequential Methods with Applications to Clinical Trials.* Boca Raton, Florida: Chapman and Hall/CRC, 2000.

[51] Jr., J. W. M., and Hammons, J. O. Delphi: A versatile methodology for conducting qualitative research. *The Review of Higher Education 18*, 4 (1995), 423–436.

[52] Kim, D., and Rhee, S. *Proceedings of the Institution of Mechanical Engineers Part B: Journal of Engineering Manufacture 175* (2004), 366–382.

[53] Kim, K., and Lin, D. Simultaneous optimization of mechanical properties of steel by maximizing exponential desirability functions. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 49*, 3 (2000), 311–325.

[54] Lachin, J. Statistical properties of randomization in clinical trials. *Controlled Clinical Trials 8* (1988), 289–311.

[55] Lachin, J. M., and Rosenberger, W. F. *Randomization in Clinical Trials Theory and Practice*, 2nd ed. ed. John Wiley & Sons, 2016.

[56] Laragh, J. H. Oral contraceptive-induced hypertension. *American Journal of Obstetrics and Gynecology 126*, 1 (1976), 141–147.

[57] Lazic, S. Ranking, selecting, and prioritising genes with desirability functions. *PeerJ: Bioinformatics and Genomics Section 3* (2015), e1444.

[58] (MACS), M. A. C. S. Macs public data set, 2017. data retrieved from `https://statepi.jhsph.edu/macs/pdt.html`, accessed 01 December 2017.

[59] (MACS), M. A. C. S. The multicenter aids cohort study (macs) is a 30-year study of the hiv-1 infection in gay and bisexual men, 2018. data retrieved from `http://aidscohortstudy.org/`, accessed 01 September 2017.

[60] Melfi, V. F., and Page, C. Estimation after adaptive allocation. *Journal of Statistical Planning and Inference 87*, 2 (2000), 353–363.

[61] Menon, S., and Zink, R. C. *Modern Approaches to Clinical Trials Using SAS: Classical, Adaptive, and Bayesian Methods.* SAS Institute, 2015.

[62] Mohajeri, L., and BLA. A statistical experiment design approach for optimizing biodegradation of weathered crude oil in coastal sediments. *Bioresource Technology 101*, 3 (2010), 893–900.

[63] Okoli, C., and Pawlowski, S. D. The delphi method as a research tool: an example, design considerations and applications. *Information & Management 42*, 1 (2004), 15–29.

[64] Pandey, R., and Panda, S. Optimization of bone drilling using taguchi methodology coupled with fuzzy based desirability function approach. *Journal of Intelligent Manufacturing 26*, 6 (2015), 1121–1129.

[65] Parzen, M., and Lipsitz, S. Dear kbg. does clustering affect the usual test statistics of no treatment effect in a randomized trial? *Biometrical Journal 40* (1998), 385–402.

[66] Pasandideh, S. H. R., and Niaki, S. T. A. Multi-response simulation optimization using genetic algorithm within desirability function framework. *Applied Mathematics and Computation 175* (2006), 366–382.

[67] Pizzaro, C., González-Sáiz, J. M., and Garrido-Vidal, D. Kinetic modelling of acetic fermentation in an industrial process by genetic algorithms with a desirability function. *Journal of Chemometrics 17*, 8-9 (2003), 453–462.

[68] Pope, J. E., Bellamy, N., Reisbold, J. R., Baron, M., Ellman, M., Carette, S., Smith, C. D., Chalmers, I. M., Hong, P., O'Hanlon, D., Kaminska, E., Markland, J., Sibley, J., Catoggio, L., and Furst, D. E. A randomized, controlled trial of methotrexate versus placebo in early diffuse scleroderma. *Arthritis & Rheumatism 44*, 6 (2001), 1351–1358.

[69] Proschan, M., and Follmann, D. Cluster without fluster: The effect of correlated outcomes on inference in randomized clinical trials. *Statistics in Medicine 27* (2008), 795–809.

[70] Rand. Rand corporation. delphi method. URL `https://www.rand.org/topics/delphi-method.html`, accessed 17 November 2017.

[71] Rao, C. *Linear Statistical Inference.* John Wiley & Sons, 1973.

[72] Rosenberger, W., Stallard, N., Ivanova, A., Harper, C., and Ricks, M. Optimal adaptive designs for binary response trials. *Biometrics 57* (2001), 909–913.

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

[74] Schildcrout, J. S., and Heagerty, P. J. Regression analysis of longitudinal binary data with time-dependent environmental covariates: bias and efficiency. *Biostatistics 6*, 4 (2005), 633–652.

[75] Schindler, D., and Hilgers, R.-D. Selecting an appropriate randomization procedure for a small population group trial on the basis of a linked optimization criterion. URL `https://www.ideal.rwth-aachen.de/wp-content/uploads/2014/02/Schindler_IWS.pdf`.

[76] Schulz, K., Chalmers, I., Hayes, R., and Altman, D. Empirical evidence of bias. dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA 273*, 5 (1995), 408–412.

[77] Su, H., and Zhu, C. *Application of Entropy Weight Coefficient Method in Evaluation of Soil Fertility. In: Qian Z., Cao L., Su W., Wang T., Yang H. (eds) Recent Advances in Computer Science and Information Engineering. Lecture Notes in Electrical Engineering, vol 126.* Springer, Berlin, Heidelberg, 2012.

[78] Suleman, S., Zeleke, G., Deti, H., Mekonnen, Z., Duchateau, L., Levecke, B., Vercuysse, J., D'Hondt, M., Wyendaele, E., and De Spiegeleer, B. Quality of medicines commonly used in the treatment of soil transmitted helminths and giardia in ethiopia: a nationwide survey. *PLoS Neglected Tropical Disease 8*, 12 (2014), e3345.

[79] Sverdlov, O., and Rosenberger, W. F. On recent advances in optimal allocation designs in clinical trials. *Journal of Statistical Theory and Practice 7*, 4 (2013).

[80] Taguchi, G. *Introduction to Quality Engineering.* Asian Productivity Organization, 1990.

[81] Taves, D. The use of minimization in clinical trials. *Contemporary Clinical Trials 31*, 2 (2010), 180–184.

[82] Torgerson, D., and Roberts, C. Understanding controlled trials. randomisation methods: concealment. *BMJ 319*, 7206 (1999), 756–376.

[83] Torgerson, D. J. What is zelen's design? *BMJ 316* (1998), 600–606.

[84] Torgerson, D. J. Contamination in trials: is cluster randomisation the answer? *BMJ 322* (2001), 355–357.

[85] Trautmann, H., and Weihs, C. On the distribution of the desirability index using harrington's desirability function. *Metrika 63* (2005), 207–213.

[86] Van Dorpe, S., Adrians, A., and Vermeire, S. Desirability function combining metabolic stability and functionality of peptides. *Journal of Peptide Science 17* (2011), 398–404.

[87] Villar, S. S., Bowden, J., and Wason, J. Response-adaptive designs for binary responses: how to offer patient benefit while being robust to time trends? *Pharmaceutical Statistics 17* (2017), 182–197.

[88] Wang, L., Chen, Y., and Zhu, H. Implementing optimal allocation in clinical trials with multiple endpoints. *Journal of Statistical Planning and Inference 182* (2016), 88–99.

[89] Wei, L. A class of designs for sequential clinical trials. *Journal of the American Statistical Association 72* (1977), 383–386.

[90] Wei, L. An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association 73* (1978), 559–563.

[91] Wong, W. Assessing disease progression using a composite endpoint. *Statistical Methods in Medical Research 16*, 1 (2007), 31–49.

[92] Wu, C. J., and Hamada, M. S. *Experiments Planning, Analysis, and Optimation*, 1st ed. ed. Wiley Series in Probability and Statistics, 2000.

[93] Wu, F. Optimization of correlated multiple quality characteristics using desirability function. *Quality Engineering 17*, 1 (2004), 119–126.

[94] Wu, S., Wong, W. K., and Crespi, C. M. Maximin optimal designs for cluster randomized trials. *Biometrics 73*, 3 (2017), 9160–926.

[95] Zhang, L., Ma, Y., and Ouyang, L. Application of the modified genetic algorithm to multi-response robust design based on the entropy weight and the desirability function. IEEE 2013 Sixth International Symposium on Computational Intelligence.

[96] Zhang, L., and Rosenberger, W. F. Response-adaptive randomization for clinical trials with continuous outcomes. *Biometrics 62* (2006), 562–569.

[97] Zhou, S., and Wang, J. Multi-response robust design based on improved desirability function. IEEE 2015 International Conference on Grey Systems and Intelligent Services (GSIS).